

# Data Mining 1

## Chapitre 5 : Analyse Factorielle Discriminante (AFD)

### Principe :

Ce document constitue des notes de cours illustrées sur un jeu de données. Chaque concept est complété par un exemple qui contient des commentaires de sorties obtenues avec le logiciel SAS et données en annexe. Ces sorties sont repérées par des numéros (AFD1, AFD2,...,AFD14). Le code SAS permettant d'obtenir les résultats est aussi fourni en annexe. En fin de cours se trouvent les sorties R sur l'exemple du cours avec les principales fonctions ainsi que des commentaires.

## 1 Généralités

### 1.1 Applications

Credit-scoring (prévision du comportement des demandeurs de crédit, diagnostic automatique, contrôle de qualité, prévision de risques (en météorologie, prévision des avalanches à partir de variables liées à l'atmosphère et à la neige),...)

### 1.2 Données

On suppose que l'on dispose de  $n$  observations décrites par  $p$  variables numériques et regroupées en  $k$  **classes ou groupes**. Ces classes sont représentées par une variable nominale à  $k$  modalités. On travaille donc sur un tableau de type individus-variables de taille  $n \times (p + 1)$ .

**Exemple 1** *Dans un cabinet de recrutement, on considère des candidats ( $n = 10$  observations) auxquels on affecte des notes ( $p = 3$  variables) selon certains critère : DIP (diplôme, de 1 pour un BTS à 5 pour une grande école d'ingénieur), TEST (tests et entretiens, de 1 à 5), EXP (expérience antérieure de 1 à 5) (Cf. AFD 1). D'autre part, on dispose pour ces personnes du résultat de leur candidature (a donné satisfaction ou pas). Ce résultat est codé par la variable nominale RES à 2 modalités (0 ou 1). On est donc en présence de  $k = 2$  classes ou sortes d'individus (Cf. AFD 2 et 3).*

**Notations :** on note  $X^1, X^2, \dots, X^p$  les  $p$  variables numériques explicatives et  $Y$  la variable nominale à expliquer.

### 1.3 Objectifs

L'analyse discriminante se propose, dans un premier temps, de séparer au mieux les  $k$  classes à l'aide des  $p$  variables considérées comme explicatives. Pour ce faire, l'AFD va constituer de nouvelles variables dites discriminantes. Ces var. conduisent à des représentations graphiques qui séparent au mieux le groupes uniquement à partir des  $p$  var. quantitatives (aspect descriptif).

Dans un deuxième temps, l'AFD cherche à résoudre le problème de l'affection d'un individu (caractérisé par ces  $p$  variables quantitatives) à l'une des  $k$  classes (aspect décisionnel). C'est un problème de classement (et non de classification).

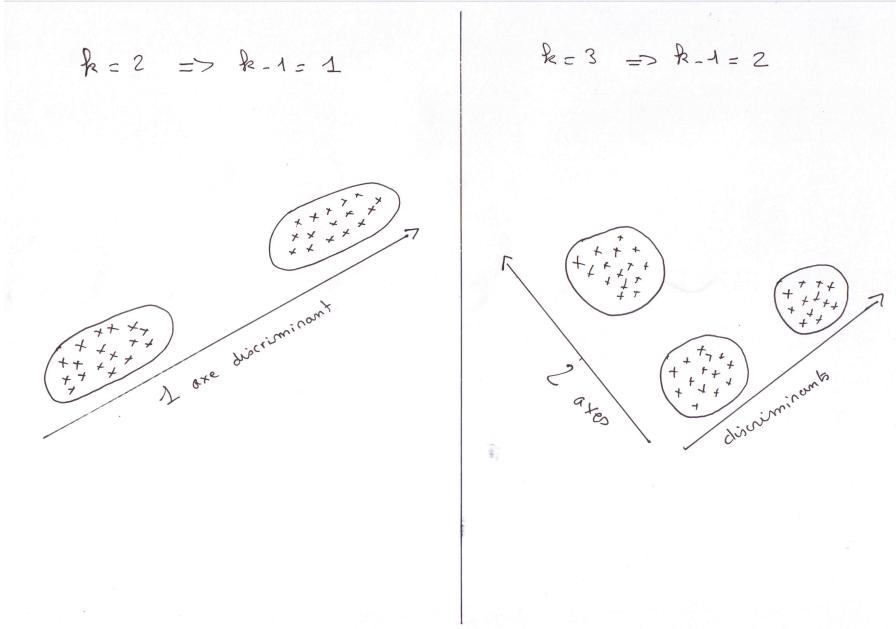
- Exemple 2**
- *Premier objectif : expliquer la variable RES par les 3 notes obtenues par les candidats et obtenir une représentation graphique qui distingue les candidats satisfaisants des autres uniquement à partir des 3 notes (Cf. AFD 9-12).*
  - *Deuxième objectif : pouvoir dire si un nouveau candidat va donner satisfaction ou pas au vu de ses 3 notes (sans disposer pour ce candidat de la variable RES).*

**Remarque 1** On est comme en régression dans le cadre d'un modèle explicatif (avec une variable à expliquer qualitative).

### 1.4 Fonctions ou variables linéaires discriminantes

Une fonction linéaire discriminante est une combinaison linéaire des  $p$  variables quantitatives qui permet de séparer au mieux les  $k$  classes.

De façon générale, remarquons qu'il faut au plus  $k - 1$  fonctions discriminantes pour séparer  $k$  classes (Cf. graphiques et AFD 2)



**Remarque 2** En AFD, on cherche donc (comme en ACP) à construire de nouvelles variables combinaisons linéaires des variables initiales (composantes principales en ACP, composantes ou variables discriminantes en AFD). Mais alors que l'objectif de l'ACP est de conserver l'inertie (ou variance) totale en ACP, il s'agit en AFD de trouver des **groupes** ou classes. Précisons (de façon mathématique) cet objectif.

## 2 AFD

### 2.1 Variance inter-classes et variance intra-classes

Au lieu de calculer les moyennes et matrices de variance sur l'ensemble de la population, on les calcule pour chaque classe (c-a-d pour chaque modalité de  $Y$ ).

**Notations :** on note :  $n_1, n_2, \dots, n_k$  les effectifs de chaque classe et

$$\overline{X^1}^{(1)}, \overline{X^1}^{(2)}, \dots, \overline{X^1}^{(k)}$$

les moyennes de la variable  $X^1$ ,

$\vdots$

$$\overline{X^p}^{(1)}, \overline{X^p}^{(2)}, \dots, \overline{X^p}^{(k)}$$

les moyennes de la variable  $X^p$  et

$$\bar{X}^{(1)}, \bar{X}^{(2)}, \dots, \bar{X}^{(k)}$$

le vecteur de moyennes pour chaque classe.

**Exemple 3** Pour la variable DIP, la moyenne globale est :

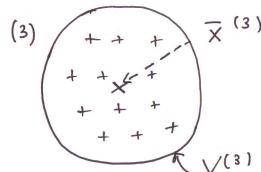
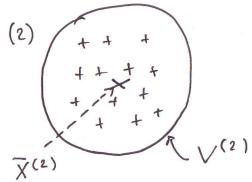
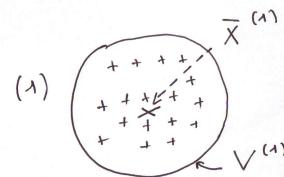
$$\overline{DIP} = (1 + 2 + 1 + 2 + \dots + 5)/10 = 23/10 = 2.3$$

Par classe (RES=0, RES=1), on a :

$$(RES = 0) \quad \overline{DIP}^{(1)} = (1 + 2 + 2 + 1 + 1)/5 = 7/5 = 1.4$$

$$(RES = 1) \quad \overline{DIP}^{(2)} = (1 + 4 + 3 + 3 + 5)/5 = 16/5 = 3.2$$

De la même façon, on calcule des matrices de variances pour chaque classe que l'on note  $V^{(j)}$ ,  $j = 1, \dots, k$ . (Cf. graphique avec  $p = 2$  variables et  $k = 3$  classes.)



On connaît la matrice de variance *totale* :

$$V = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}) (X_i - \bar{X})'$$

On définit maintenant 2 nouvelles matrices de variance :

**Définition 1** – *la matrice de variance intra-classe* :

$$W = \frac{1}{n} \sum_{j=1}^k n_j V^{(j)}$$

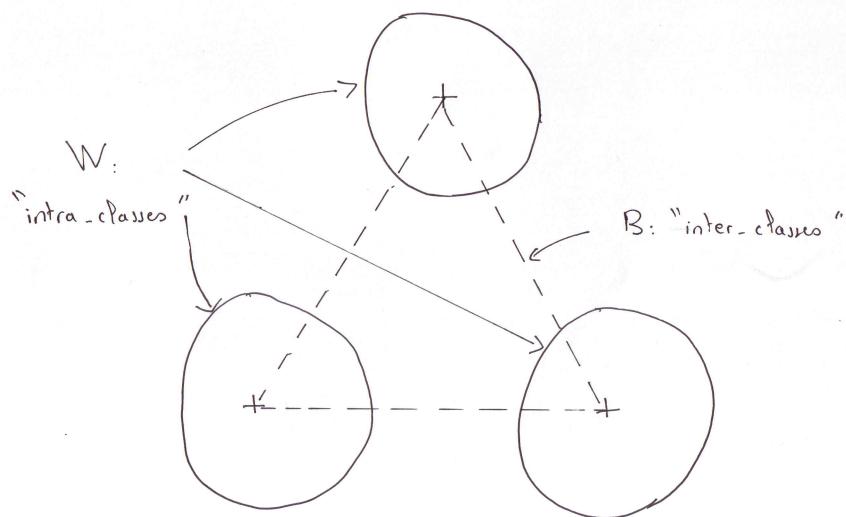
(moyenne des matrices de var. de chaque classe,  $W$  pour within)

– *la matrice de variance inter-classe* :

$$B = \frac{1}{n} \sum_{j=1}^k n_j (\bar{X}^{(j)} - \bar{X}) (\bar{X}^{(j)} - \bar{X})$$

(matrice de variance des  $k$  moyennes,  $B$  pour between)

(Cf. graphique)



**Proposition 1** *On a la relation suivante :*

$$V = W + B$$

*(équation d'analyse de la variance)*

**Intuition :** en AFD, on cherche des variables (ou axes) telles que sur ces axes, la matrice de variance inter-classes soit la plus grande possible tandis que la matrice intra est la plus petite possible (alors qu'en ACP on cherche la variance totale la plus grande possible).

**Remarque 3** *on peut noter le parallèle entre l'AFD et la régression linéaire. En régression, on a :  $SCT=SCE+SCR$  et on cherche  $SCR$  minimale et donc  $SCE$  maximale ( $SCT$  fixée). On a  $SCT \rightarrow V$ ,  $SCE \rightarrow B$  et  $SCR \rightarrow W$ .*

## 2.2 Distance de Mahalanobis

**Rappel :** si on a 2 vecteurs  $u$  et  $v$  de  $\mathbb{R}^p$ , la distance usuelle entre ces 2 vecteurs est donnée par :

$$d^2(u, v) = \sum_{i=1}^p (u_i - v_i)^2$$

soit matriciellement,

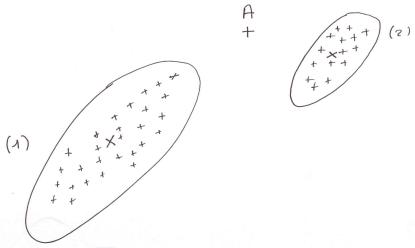
$$d^2(u, v) = (u - v)'(u - v)$$

**Définition 2** *on appelle distance de Mahalanobis la distance définie par :*

$$d_W^2(u, v) = (u - v)'W^{-1}(u - v)$$

où  $W$  désigne la matrice de variances intra-classes.

**Explication graphique :**



Même si le point A est plus proche du centre du groupe (2), compte tenu de l’“élongation” des groupes, le point A fait plutôt partie du groupe (1). Or, cette élongation est justement mesurée par la matrice  $W$  d'où l'utilisation en AFD de la distance de Mahalanobis (comme la distance du  $\chi^2$  en AFC).

**Exemple 4** SAS utilise la distance de Mahalanobis dans sa procédure d'AFD. En particulier, il donne les distances de Mahalanobis au carré entre les moyennes des groupes (Cf. AFD 4).

## 2.3 Étapes de l'analyse

### 2.3.1 Calculs des variables discriminantes (calculs de type ACP)

Ces variables sont notées CAN1,CAN2,... comme dans SAS. On cherche les valeurs propres et vecteurs propres de la matrice :

$$W^{-1} B$$

(au lieu de  $R$  en ACP). On rappelle que l'on en a  $k - 1$ .

**Proposition 2** Les variables discriminantes ne sont pas corrélées.

**Question :** où intervient la variable  $Y$  puisque dans les formules de  $B$  et de  $W$  n'apparaissent que les var. quantitatives  $X^1, \dots, X^p$ ? En fait faux ! Pour calculer  $B$  et  $W$ , on doit connaître l'affectation des ind. aux groupes. et donc  $Y$  intervient à ce niveau.

SAS fournit les valeurs propres (Cf. AFD 6) et les vecteurs propres (appelés *canonical coefficients*). En fait, SAS donne 3 sortes de vecteurs propres selon différentes normalisations :

Total-Sample Standardized Canonical Coefficients

Pooled Within-Class Standardized Canonical Coefficients

Raw Canonical Coefficients

On utilisera les coefficients directs (Raw Canonical Coefficients) pour calculer les var. discriminantes directement à partir des var. centrées. En effet, comme en ACP pour les  $p$  comp. ppales, on calcule les  $k - 1$  variables discriminantes à partir des données et des vecteurs propres (en ACP,  $c = x \times v$ ).

**Exemple 5** Voir le tableau AFD8. On a :

Raw Canonical Coefficients

	CAN1
DIP	1.237566934
TEST	0.632047850
EXP	2.178011688

On centre simplement les données et on multiplie chaque obs. par le vecteur associé. On obtient :

$$\begin{aligned} CANI &= 1.238 \times (DIP - \overline{DIP}) + 0.632 \times (TEST - \overline{TEST}) + 2.18 \times (EXP - \overline{EXP}) \\ &= \begin{pmatrix} -0.970 \\ -3.175 \\ 0.576 \\ -0.996 \\ -1.602 \\ 1.479 \\ 0.873 \\ -1.320 \\ 1.787 \\ 3.348 \end{pmatrix} \end{aligned}$$

avec par exemple pour la première obs. :

$$\overline{DIP} = 2.3, \overline{TEST} = 3.3, \overline{EXP} = 4.2,$$

$$1.238 \times (1 - 2.3) + 0.632 \times (5 - 3.3) + 2.178 \times (4 - 4.2) = -0.97$$

### 2.3.2 Qualité globale de la discrimination

L'AFD parvient plus ou moins bien à discriminer les groupes d'individus à partir des  $p$  variables fournies. Pour savoir si globalement l'analyse a bien fonctionné, on utilise le **coefficient de corrélation canonique** au carré noté *CanRsq* par SAS. Il correspond au quotient de la somme des carrés inter-classes de la variable discriminante sur la somme des carrés totale. Il est équivalent au  $R^2$  en régression (SCE/SCT). Il est compris entre 0 et 1 et s'interprète comme le  $R^2$  en régression. Plus il est proche de 1, meilleure est la discrimination.

**Exemple 6** Cf. AFD 5 : Squared Canonical Correlation=76,5% donc, bonne discrimination.

### 2.3.3 Choix de la dimension

Ce choix se pose dès que l'on a plus de 2 variables discriminantes (car sinon, on peut faire le dessin sur une feuille de papier) c'est-à-dire au moins 4 groupes. Dans l'exemple, pas de pb car 2 classes uniquement donc une seule variable discriminante que l'on garde bien évidemment. Mais si 4 groupes ou plus, on doit utiliser un critère de choix en sachant que les variables discriminantes sont de moins en moins intéressantes. Comme en ACP, le choix se base sur les valeurs propres qui sont décroissantes (attention, en AFD, les valeurs propres ne sont pas égales aux variances des nouvelles variables discriminantes).

On a la formule :  $\lambda_j = \text{somme des carrés-inter(CANj)} / \text{somme des carrés-intra(CANj)}$ .

CANj est donc d'autant plus discriminante que  $\lambda_j$  est élevé. SAS donne le ratio de de chaque valeur propre sur le total des valeurs propres, ainsi que le cumul des ratios. (Cf. AFD 6, colonne *Cumulative*). On utilise ce cumul comme en ACP.

### 2.3.4 Interprétation des facteurs discriminants

**Coefficients** On peut utiliser les coefficients qui interviennent dans le calcul les variables discriminantes, c'est-à-dire les vecteurs propres (raw canonical coefficients).

**Exemple 7** On a :

$$CAN1 = 1.238 \times (DIP - \overline{DIP}) + 0.632 \times (TEST - \overline{TEST}) + 2.18 \times (EXP - \overline{EXP})$$

Le facteur le plus important est celui d'*EXP* mais aussi celui de *DIP*, *TEST* joue un moindre rôle.

**Corrélations et graphique des corrélations** On peut calculer les corrélations entre les variables initiales et les variables discriminantes (comme en ACP). SAS les calcule dans un tableau appelé *Total Canonical Structure* (Cf. AFD 7). On interprète alors ces corrélations ou bien directement, ou bien à partir d'un graphique avec cercle des corrélations.

**Exemple 8** Cf. AFD 7 : la variable discriminante CAN1 est corrélée positivement essentiellement avec la var. DIP mais aussi moyennement avec EXP (et très peu avec TEST). Ici, graphique inutile car une seule var. discriminante et peu de var. initiales mais Cf. TD.

**Remarque 4** Contrairement à l'ACP, il n'existe pas de relation simple entre les vecteurs propres et les corrélations (du type  $r = \sqrt{\lambda} \times v$ ). Il est donc intéressant de commenter les 2.

### 2.3.5 Graphique des observations sur les axes discriminants

**Construction** On utilise les valeurs des obs. sur les axes discriminants (CAN1, CAN2, ...) pour les représenter plan par plan :

(CAN1,CAN2), (CAN1,CAN3),...

(comme en ACP avec les plans principaux).

**Exemple 9** Cf. AFD 14 avec la variable RES car sinon un seul axe de représentation.

**Interprétation** On peut interpréter tous les individus. On repère les groupes d'individus. On précise quels axes discriminent quels groupes. On repère aussi éventuellement les zones de chevauchement entre groupes.

**Exemple 10** Cf. AFD 14 : L'axe CAN1 discrimine parfaitement les candidats satisfaisants (**scores positifs**) des autres (**scores négatifs**).

**Remarque 5** Dans d'autres cas, la discrimination n'est pas aussi bonne et on peut avoir des valeurs de variables discriminantes (ou **scores**) correspondant à une zone d'indécision où les individus ne sont pas correctement classés.

### 3 Règles d'affectation (ou de classement)

Il s'agit de répondre aux 2ème objectif de l'AFD. En effet, une fois trouvées les fonctions discriminantes qui séparent au mieux les individus répartis en  $k$  classes, on veut trouver la classe d'affectation d'un nouvel individu, pour lequel on connaît les valeurs des variables

$$X^1, \dots, X^p.$$

**Exemple 11** *L'objectif est de pouvoir dire si un nouveau candidat va donner satisfaction ou pas au vu de ses 3 notes (sans disposer pour ce candidat de la variable RES). Si le résultat obtenu est 0 (non), le cabinet peut ne pas proposer le candidat à l'entreprise.*

Il existe plusieurs règles d'affectation ou de classement en AFD. Nous choisissons d'en expliciter une qui est d'ordre géométrique. Toutefois, SAS utilisant par défaut une autre règle (probabiliste), nous la présenterons aussi de manière succincte.

#### 3.1 Règle géométrique de classement

Cette règle simple consiste à choisir la classe dont la moyenne est la plus proche de l'individu. “La plus proche” doit s'entendre au sens de la distance de Mahalanobis. Puisque ce sont ces distances qui sont représentées sur le graphique des variables discriminantes, il suffit de connaître les moyennes des classes (données par SAS) et les coordonnées des points ( $CAN_1, \dots, CAN_{(k-1)}$ ) pour pouvoir décider de l'affectation.

En fait, on a le résultat théorique suivant :

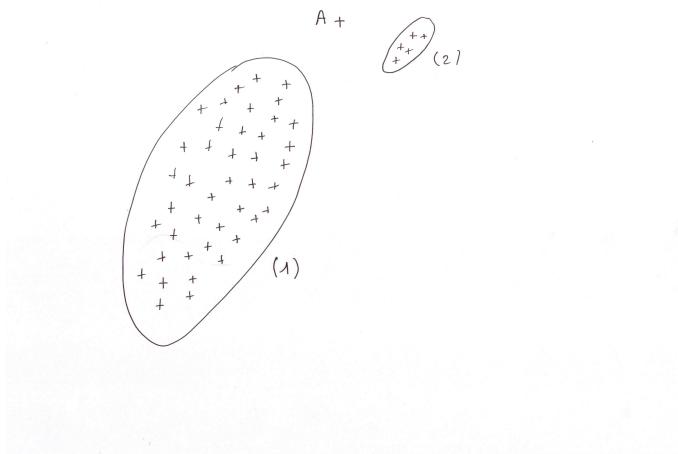
$$d_{W^{-1}}^2(X_i, X_j) = d^2(CAN_i, CAN_j)$$

où  $CAN = ((CAN_1, \dots, CAN_{(k-1)}))$

**Exemple 12** *Voir le tableau AFD9. Puisque les centres de classes sont équidistants de 0 (toujours le cas si il n'y a que 2 classes avec le même effectif dans chaque classe), on affecte à la classe satisfaisants, les individus qui obtiennent un score ou une valeur pour la variable discriminante positive et à la classe non-satisfaisants, les individus qui ont des scores négatifs. Seul la valeur 0 est à égale distance des 2 moyennes et pose un problème d'affectation (Cf. graphique).*

**Remarque 6** *La règle d'affectation permet de choisir un groupe pour un nouvel individu. Cette règle peut parfois conduire à des erreurs de classement en particulier lorsque les 2 groupes ne sont pas clairement séparés (Cf. TD).*

Le défaut majeur de la méthode est qu'elle conduit à des erreurs si les groupes se trouvent représentés dans des proportions très différentes (Cf. graphique) ce qui est le cas dans la prévision d'evts rares par exemple.



### 3.2 Règles probabilistes

On considère les distributions de probabilités des observations à l'intérieur de chaque classe (lois Normales en général). On tient éventuellement compte à ce niveau de probabilités *à priori* d'être dans un groupe (permet de résoudre le pb des événements rares). Puis, on calcule pour chaque individu des probabilités *à posteriori* de se trouver dans chacune des classes. On affecte alors les individus au groupe correspondant à la plus forte probabilité .

### 3.3 Mesure d'efficacité des règles de classement

Pour avoir une idée sur l'efficacité des règles de classement proposée, on applique la **méthode de resubstitution**, c'est-à-dire que l'on classe les individus du tableau pour lesquels on connaît déjà le groupe.

SAS donne par défaut les résultats d'affectation obtenus par la règle probabiliste pour les *individus de départ* dans un tableau commenté en TD. Or, pour ces ind., on connaît le groupe et donc, on peut repérer d'éventuelles erreurs.

- SAS calcule le taux d'erreur : nombre d'individus mal classés / nombre d'individus.
- SAS précise le type d'erreurs dans un tableau commenté en TD.

**Exemple 13** Voir le tableau AFD10. Les 2 règles d'affectation donnent des résultats identiques sur cet exemple puisque les 10 individus sont bien classés.

**Remarque 7** – Même si SAS ne fournit pas les résultats pour la règle d'affectation géométrique, il est évidemment facile de les calculer.

- Les taux sont calculés sur les observations qui ont contribué à créer les variables discriminantes. Ils sont donc optimistes par rapport à des taux calculés sur de nouveaux individus (pour lesquels on connaîttrait aussi le groupe).
- Autre méthode d'évaluation d'une règle de décision : échantillon test. Cette méthode consiste à partager l'échantillon en deux parties : une partie de l'ordre de 80% servant d'échantillon d'apprentissage de la règle de décision et l'autre servant à la tester (option testdata de la proc discrim de SAS). Cette méthode est plus fiable que la méthode de resubstitution mais nécessite un échantillon plus important.

## 4 Conclusion

Justification de l'AFD :

- tableau de type individus / variables avec des individus regroupés en classes,
- recherche d'une discrimination entre les groupes,
- affectation à l'un des groupes d'un nouvel individu caractérisé par les mêmes variables que les individus de départ.

**Remarque 8** L'AFD telle qu'elle est présentée dans ce chapitre est appelée AFD canonique par SAS.

## Annexe 1 : sorties SAS

Tableau AFD1

	Obs	Id	Dip	Test	Exp	Res
	1	A	1	5	4	0
	2	B	2	3	3	0
	3	C	1	4	5	1
	4	D	2	3	4	0
	5	E	1	4	4	0
	6	F	4	3	4	1
	7	G	3	4	4	1
	8	H	1	1	5	0
	9	I	3	2	5	1
	10	J	5	4	4	1

----- Res=0 -----					
Variable	Nb	Moyenne	Écart-type	Minimum	Maximum
Dip	5	1.4000000	0.5477226	1.0000000	2.0000000
Test	5	3.2000000	1.4832397	1.0000000	5.0000000
Exp	5	4.0000000	0.7071068	3.0000000	5.0000000

----- Res=1 -----					
Variable	Nb	Moyenne	Écart-type	Minimum	Maximum
Dip	5	3.2000000	1.4832397	1.0000000	5.0000000
Test	5	3.4000000	0.8944272	2.0000000	4.0000000
Exp	5	4.4000000	0.5477226	4.0000000	5.0000000

Tableau AFD2

### The DISCRIM Procedure

Observations	10	DF Total	9
Variables	3	DF Within Classes	8
Classes	2	DF Between Classes	1

Tableau AFD3

Class Level Information					
Res	Nom de variable	Fréquence	Pondération	Proportion	Probabilité a priori
0	_0	5	5.0000	0.500000	0.500000
1	_1	5	5.0000	0.500000	0.500000

Tableau AFD4

## Pairwise Generalized Squared Distances Between Groups

$$D_{ij}^2 = \frac{1}{n_i n_j} (X_i - \bar{X}_j)' \text{COV}^{-1} (X_i - \bar{X}_j)$$

## Generalized Squared Distance to Res

De	Res	0	1
0		0	10.40214
1		10.40214	0

Tableau AFD5

Corrélation canonique	Erreur std	Corrélation canonique
Corrélation canonique	ajustée	approchée au carré
1 0.874496	0.856034	0.078419 0.764743

Tableau AFD6

Valeurs propres de Inv(E) *H= CanRsq/(1-CanRsq)
Valeur propre
Différence
3.2507
Proportion Cumulée
1.0000 1.0000

Tableau AFD7

Structure canonique totale

Variable	Can1
Dip	0.764972
Test	0.103956
Exp	0.381172

Tableau AFD8

Coefficients canoniques bruts

Variable	Can1
Dip	1.237566934
Test	0.632047850
Exp	2.178011688

Tableau AFD9

Moyennes de classes sur les variables canoniques

Res	Can1
0	-1.612617363
1	1.612617363

Tableau AFD10

Generalized Squared Distance Function

$$D_j^2 = (X - \bar{X}_j)' \text{COV}^{-1} (X - \bar{X}_j)$$

Posterior Probability of Membership in Each Res

$$\Pr(j|X) = \frac{\exp(-.5 D_j^2)}{\sum_k \exp(-.5 D_k^2)}$$

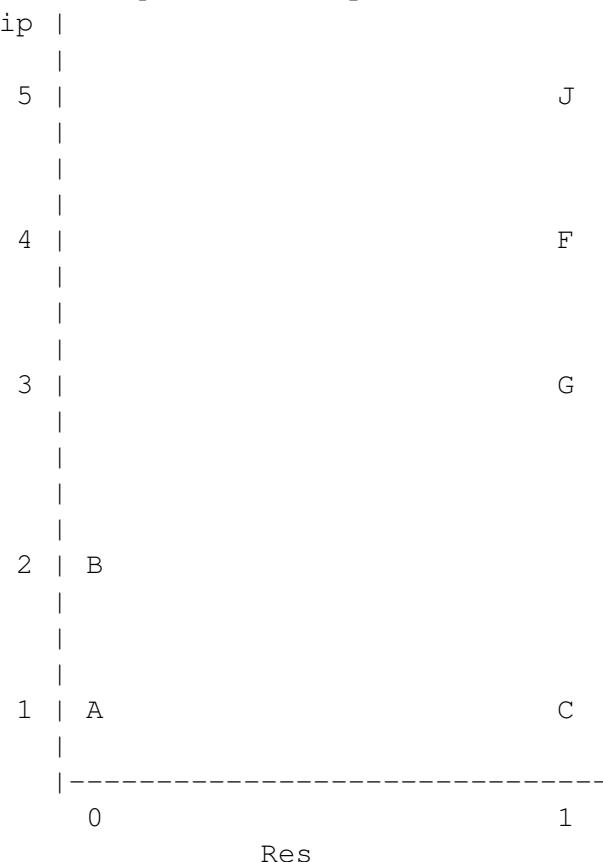
Number of Observations and Percent Classified into Res			
De Res	0	1	Total
0	5	0	5
	100.00	0.00	100.00
1	0	5	5
	0.00	100.00	100.00
Total	5	5	10
	50.00	50.00	100.00
Priors	0.5	0.5	

Error Count Estimates for Res			
	0	1	Total
Rate	0.0000	0.0000	0.0000
Priors	0.5000	0.5000	

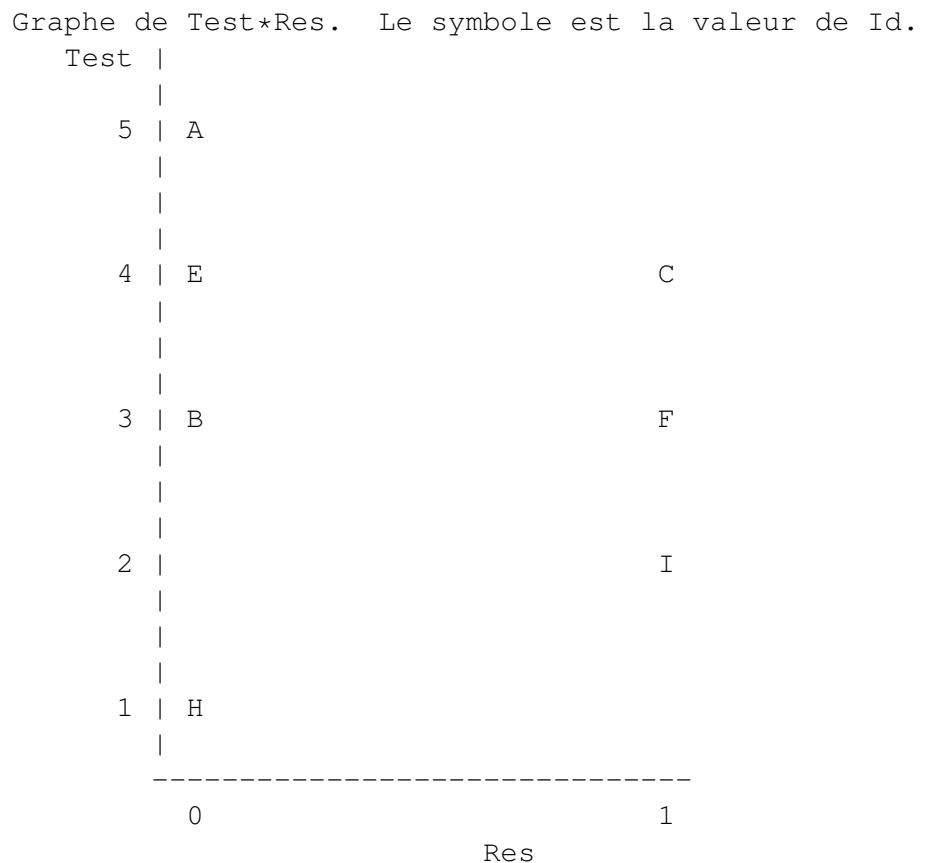
Graphe AFD11

Graphe de Dip\*Res. Le symbole est la valeur de Id.



NOTE : 4 obs cachée(s).

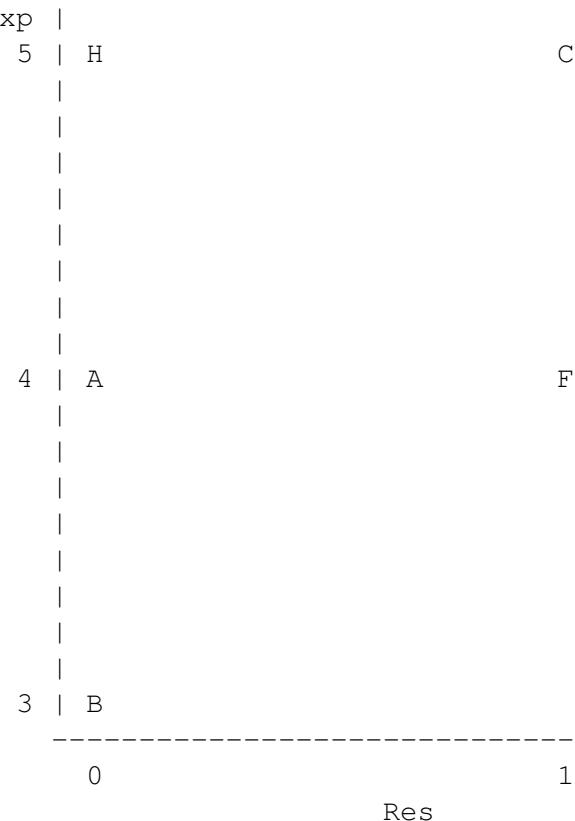
Graphe AFD12



NOTE : 3 obs cachée(s) .

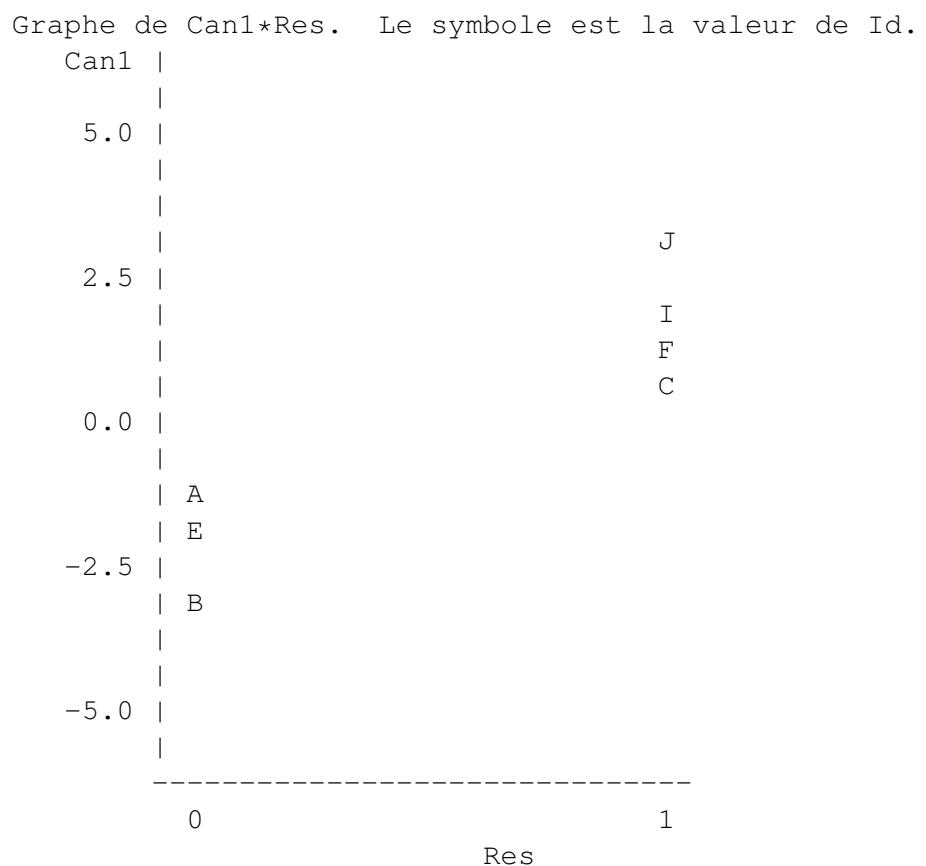
Graph AFD13

Graphe de Exp\*Res. Le symbole est la valeur de Id.



NOTE : 5 obs cachée(s).

Graph AFD14



NOTE : 3 obs cachée(s).

## Annexe 2 : code SAS

```
/* lecture des donnees */
data recrut;
infile 'F:\m1\2008\recrutement.txt';
input Id$ Dip Test Exp Res;
run;
proc print;
run;
proc sort data=recrut;
by Res;
run;
proc means data=recrut;
by Res;
run;
/* afd */
proc discrim data=recrut out=resafdrecrut canonical;
var Dip Test Exp;
class res;
run;

proc plot;
plot Dip*Res=Id /vpos=20 hpos=40;
run;

proc plot;
plot Test*Res=Id /vpos=20 hpos=40;
run;

proc plot;
plot Exp*Res=Id /vpos=20 hpos=40;
run;

proc plot data=resafdrecrut;
plot can1*Res=Id /vpos=20 hpos=40;
run;
```

### Annexe 3 : sorties R (fonction lda) du package MASS

```
# charger la table recrutement
attach(recrutement)
library(MASS)
resu_afd=lda(Res~Dip+Test+Exp)
resu_afd
Call:
lda(Res ~ Dip + Test + Exp)

Prior probabilities of groups:
 0   1
0.5 0.5

Group means: moyennes des variables par groupe
  Dip Test Exp
0 1.4  3.2 4.0
1 3.2  3.4 4.4

Coefficients of linear discriminants: coefficients pour calculer
la variable discriminante à partir des variables centrées
  LD1
Dip  1.2375669
Test  0.6320478
Exp   2.1780117

#Pour avoir d'autres sorties
> names(resu_afd)
[1] "prior"    "counts"   "means"    "scaling"   "lev"       "svd"       "N"
[8] "call"     "terms"    "xlevels"

> resu_afd$svd
[1] 5.099544 " voir ci-dessous

#Calcul de la variable discriminante (à ne pas faire en pratique, juste pour comprendre)
Dipcentre=Dip-mean(Dip)
Testcentre=Test-mean(Test)
> Testcentre=Test-mean(Test)
> Expcentre=Exp-mean(Exp)
```

```

> CAN1=1.2375669*Dipcentre+0.6320478*Testcentre+2.1780117*Expcentre
> CAN1
[1] -0.9699581 -3.1744984  0.5760058 -0.9964868 -1.6020059  1.4786470
[7]  0.8731279 -1.3201376  1.7870440  3.3482617

# la fonction predict() fournit directement CAN1
> predict(resu_afd)$x
      LD1
1 -0.9699580
2 -3.1744985
3  0.5760058
4 -0.9964868
5 -1.6020059
6  1.4786471
7  0.8731280
8 -1.3201377
9  1.7870440
10 3.3482619

# Calcul des moyennes par groupe
> tableau=cbind(CAN1,Res)
> mean(tableau[Res==0,1])
[1] -1.612617
> mean(tableau[Res==1,1])
[1] 1.612617
>
On trouve des moyennes opposées car 2 groupes
de même effectif

#Interprétation de CAN1
> cor(CAN1,Dipcentre)
[1] 0.7649719
> cor(CAN1,Testcentre)
[1] 0.103956
> cor(CAN1,Expcentre)
[1] 0.3811721

CAN1 est surtout corrélée avec le diplôme

```

```
#Régression linéaire (ANOVA à 1 facteur) de CAN1 sur la variable
qualitative Res
pour savoir si Res a un effet significatif sur CAN1 c'est-à-dire
si CAN1 discrimine significativement Res.
> summary(lm(CAN1~as.factor(Res)))
```

Call:  
`lm(formula = CAN1 ~ as.factor(Res))`

Residuals:

Min	1Q	Median	3Q	Max
-1.56188	-0.58811	0.09252	0.53522	1.73564

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-1.6126	0.4472	-3.606	0.00692 **
as.factor(Res) [T.1]	3.2252	0.6325	5.100	0.00093 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1 on 8 degrees of freedom  
Multiple R-squared: 0.7647, Adjusted R-squared: 0.7353  
F-statistic: 26.01 on 1 and 8 DF, p-value: 0.0009304

On retrouve le R<sup>2</sup> = 76,47 % (canonical R squared fourni par SAS)  
d'où bonne discrimination

On a aussi le test H<sub>0</sub> : absence d'effet de Res sur CAN1,  
on a F=26.01 = carré-moyen inter/carré moyen intra  
= (somme carrés inter/k-1)/ (somme carrés intra/n-k)  
=(svd<sup>2</sup>) et p=0.0009304<5%.

Donc CAN1 discrimine significativement Res.

```
#Calcul des probabilités à posteriori
proba=predict(resu_afd)$posterior
```

\$posterior

	0	1
1	9.580468e-01	4.195319e-02
2	9.999642e-01	3.576508e-05
3	1.349651e-01	8.650349e-01
4	9.613542e-01	3.864584e-02
5	9.943298e-01	5.670198e-03
6	8.417544e-03	9.915825e-01
7	5.646339e-02	9.435366e-01
8	9.860435e-01	1.395652e-02
9	3.129847e-03	9.968702e-01
10	2.042092e-05	9.999796e-01

```

## validation de l'AFD
# A l'aide des probabilités (comparées à 0,5 : par exemple, l'individu 1 a
une proba d'être dans le groupe 0 égale à
0.958 > 0.5 donc il est classé dans le groupe 1 par la méthode),
un classement dans les groupes 0 ou 1 est prédit :
predict(resu_afd)$class
[1] 0 0 1 0 0 1 1 0 1 1
Levels: 0 1

```

On compare le groupe prédict au groupe de départ :

```
prediction=as.vector(predict(resu_afd)$class)
```

```
> table(Res,prediction)
    prediction
Res 0 1
  0 5 0
  1 0 5
```

On conclut que sur cet exemple, tous les individus sont bien classés

Graphique : si l'échantillon est important, faire plutôt des histogrammes ou des boîtes à moustaches par groupe

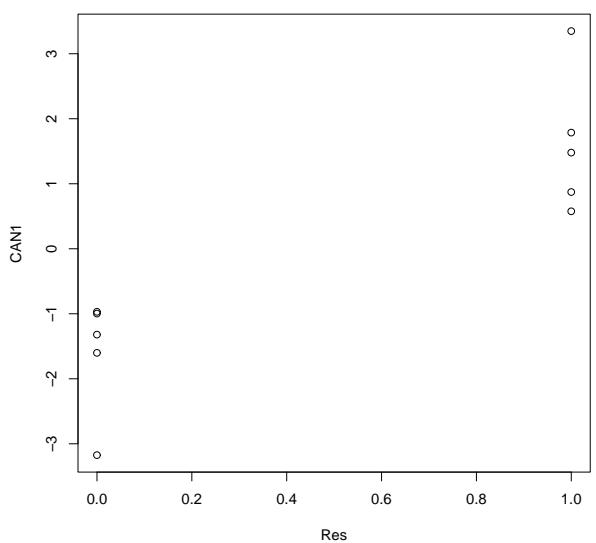


FIGURE 1 – Nuage de points de la variable discriminante par groupe