Université d'Angers **Janvier 2007**
Master 1 MIM : Analyse des données

Examen d'Analyse des données

Durée : 3 heures
Les documents ne sont pas autorisés. La calculatrice est autorisée.

Exercice I (7 points) : ACP non normée

On dispose du classement de 11 individus sur 3 matières : math, musique et français.
Le classement en math revient à numéroter les individus. Le tableau des classements selon les trois matières est le suivant :

Math	1	2	3	4	5	6	7	8	9	10	11
Musique	6	1	4	5	3	2	9	7	8	10	11
Français	2	6	5	3	4	1	8	9	7	10	11

Chaque individu est affecté du même poids.
Pour les calculs, vous pouvez utiliser les valeurs arrondies au millième.

I. EFFECTUER L'ACP NON NORMEE DU TABLEAU DE RANGS

1) Calculer le centre de gravité g_I du nuage des individus.

2) Calculer le tableau centré Y (centré en lignes).

3)
a. Calculer la matrice d'association V du nuage des individus *N(I)*.
b. Que représente cette matrice ?
c. Quelle est l'inertie du nuage ?

4) Recherche des axes principaux d'inertie
a. Démontrer que les trois valeurs propres sont : 25,090; 2.455; 2.455.
b. Vérifier votre réponse à l'aide du 3).
c. Retrouver par le calcul le vecteur $U_1 = \begin{pmatrix} 0.577 \\ 0.577 \\ 0.577 \end{pmatrix}$ correspondant au premier vecteur propre.

On donne $U_2 = \begin{pmatrix} 0.811 \\ -0.486 \\ -0.326 \end{pmatrix}$ $U_3 = \begin{pmatrix} -0.093 \\ -0.656 \\ 0.749 \end{pmatrix}$.

5)
a. Quelle est la contribution absolue de l'axe F_1 à l'inertie du nuage ?
b. Quel est le taux d'inertie extrait par l'axe F_1 ?
c. Quelle est la meilleure représentation à une dimension du nuage ?
d. Quelle est la meilleure représentation plane ?

II. REPRESENTATION DES INDIVIDUS

1) Compléter dans le tableau ci-dessous les composantes principales (coordonnées des individus).
2) Effectuer la représentation graphique du plan F_1-F_2.

composantes principales				qualité de représentation (/1000)				contribution (/1000)	
	F1	**F2**	**F3**		**F1**	**F2**	**F3**		**F1**
ind1				ind1	658,537	185,010	156,454	ind1	97,838
ind2				ind2	658,537	16,253	325,211	ind2	97,838
ind3	-3,464	1,137	0,841	ind3	857,143	92,300	50,557	ind3	43,484
ind4	-3,464	-0,160	-1,405	ind4	857,143	1,834	141,024	ind4	43,484
ind5	-3,464	1,297	0,564	ind5	857,143	120,152	22,705	ind5	43,484
ind6	-5,196	3,570	-1,119	ind6	658,537	310,933	30,531	ind6	97,838
ind7	3,464	-1,297	-0,564	ind7	857,143	120,152	22,705	ind7	43,484
ind8	3,464	0,160	1,405	ind8	857,143	1,834	141,024	ind8	43,484
ind9	3,464	1,137	-0,841	ind9	857,143	92,300	50,557	ind9	43,484
ind10	6,928	0,000	0,000	ind10	1000,000	0,000	0,000	ind10	173,934
ind11	8,660	0,000	0,000	ind11	1000,000	0,000	0,000	ind11	271,772

III. REPRESENTATION DES VARIABLES

1) Calculer les coordonnées de la variable math sur les différents axes et compléter le tableau.
2) Effectuer la représentation graphique dans les différents plans.
3) En quoi cette représentation illustre-t-elle le I-3) ?

coordonnées des variables				qualité de représentation			
	F1	F2	F3		F1	F2	F3
math				math	836,333	161,566	2,101
musi	2,892	-0,761	-1,028	musi	836,333	57,922	105,745
fran	2,892	-0,510	1,173	fran	836,333	26,012	137,654

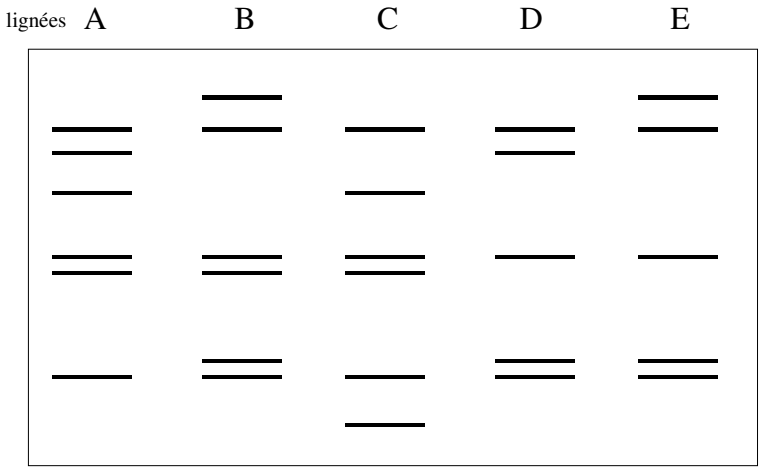
III. ANALYSER BRIEVEMENT LES RESULTATS OBTENUS

IV. INDIVIDUS SUPPLEMENTAIRES

Un auditeur libre a eu des notes qui l'auraient classé 8ème en math, 2ème en musique et 1er en français.
Situer cet individu par rapport à l'ensemble des 11 autres dans le plan F_1-F_2.

Exercice II : Classification (5 points)

Un laboratoire veut étudier la ressemblance génétique de cinq lignées de sorgho (A, B, C, D, E). Pour cela, il réalise une analyse par RFLP avec l'enzyme de restriction *Eco* RI et une sonde d'origine inconnue. Les fragments d'ADN ainsi amplifiés sont ensuite séparés par electrophorèse. Les résultats de l'electrophorèse donnent les profils suivants:



NB : Deux lignées de sorgho présentent autant de fragments d'ADN identiques que de bandes révélées à la même hauteur sur les profils de l'electrophorèse.

- Question 1 :
  Construisez la matrice de similarité entre les cinq lignées de sorgho à partir du coefficient de similarité de Dice (Sxy) défini ci-dessous.

$$Sxy = 2\frac{Nxy}{Nx + Ny}$$

Nx = nombre de bandes de la lignée X

Ny = nombre de bandes de la lignée Y

Nxy = nombre de bandes identiques entre les lignées X et Y

Similarité	A	B	C	D	E
A					
B					
C					
D					
E					

- Question 2 :

Déterminez la matrice de dissimilarité en utilisant la fonction de similitude linéaire :
Dxy = C – Sxy *où C = max Sxy*

- Question 3 :

Effectuez sur la matrice de dissimilarité la CAH en utilisant le critère du saut maximal. Votre réponse doit comporter les différentes étapes et la représentation de l'arbre hiérarchique.

- Question 4 :

Que concluez vous sur les ressemblances entre les lignées de sorgho ?

III Analyse discriminante (8 points) : Dystrophie musculaire de Duchenne

Cette maladie est causée par la modification d'un gène responsable de la fabrication de la protéine dystrophine, laquelle contribue à la force et la santé des muscles. Cette modification est ce qu'on appelle une mutation. Lorsque ce gène subit une mutation, la protéine dystrophine ne s'acquitte plus de sa fonction. Les cellules musculaires s'affaiblissent et se détruisent peu à peu. Rarissime chez les filles, cette maladie affecte surtout les garçons.

Nous disposons de quatre variables pouvant être utilisés comme prédicteurs de cette maladie : la créatine kinase, l'hemopexine, la lactate dehydrogénase et la pyruvate kinase (enzymes).

```
> tableau
    CREATKIN HEMOPEX LACTDEHY PYRUKIN CARRIER
1        5.2      8.4       1.1     17.6       1
2        2.0      7.7       1.1     20.0       1
```

CARRIER indique l'absence (1) ou la présence (2) de la maladie

Les questions qui se posent alors sont :

- Existent-ils des différences dans la concentration entre ces différentes enzymes en relation avec l'absence ou la présence de cette maladie génétique?
- Etant donné des valeurs de ces concentrations, peut-on prévoir la présence ou l'absence de la maladie ?

Les résultats sont présentés dans les deux pages suivantes.

1. **a)** Qu'appelle-t-on fonction linéaire discriminante?
b) Rappeler le critère utilisé pour déterminer les fonctions linéaires discriminantes.
c) Combien de fonctions linéaires discriminantes peut-on déterminer dans cet exemple.

2. **a)** A l'aide du résultat e., déterminer la première fonction discriminante dans cet exemple.
b) Retrouver la coordonnée F_1 de l'individu 1 dans le tableau f..

3. **a)** Quelle information apporte le test F au résultat c.?
b) Interpréter le F* au résultat d.? Que peut-on conclure?

Dans la suite, on suppose que les deux classes suivent des lois multinormales.

3. **a)** Quelle est la dimension de ces lois?
b) Donner une estimation de la moyenne du groupe 1 à l'aide du résultat c.
c) Donner une estimation de la matrice des covariances sous l'hypothèse où elle est identique dans les deux classes à l'aide du résultat a..

4. Le test de Bartlett est présenté au résultat d.. Ils portent sur l'égalité des moyennes.
a) Quelle est l'hypothèse nulle pour le test portant sur F_1.
b) Interpréter le résultat obtenu ici.

6. La qualité de classement obtenue a été calculée sur l'échantillon d'apprentissage.
a) Comment interpréter le tableau g. Préciser la signification de Sc1 Sc2 P et Er. Comment est déterminé le classement *a posteriori*?
b) Déterminer les individus mal classés dans ce tableau.
c) Déterminer le taux d'erreur du classement dans chaque population.

8. Proposer deux méthodes pour améliorer l'évaluation de la qualité du classement en AFD.

Exercice III - Résultats de l'AFD

a. $varT

	CREATKIN	HEMOPEX	LACTDEHY	PYRUKIN
CREATKIN	10921.58	272.25	1162.69	3301.56
HEMOPEX	272.25	134.72	34.68	252.49
LACTDEHY	1162.69	34.68	183.86	486.81
PYRUKIN	3301.56	252.49	486.81	3355.68

$varW

	CREATKIN	HEMOPEX	LACTDEHY	PYRUKIN
CREATKIN	8620.38	-23.66	881.20	1614.46
HEMOPEX	-23.66	96.44	-1.33	35.66
LACTDEHY	881.20	-1.33	149.77	280.44
PYRUKIN	1614.46	35.66	280.44	2118.33

$varB

	CREATKIN	HEMOPEX	LACTDEHY	PYRUKIN
CREATKIN	2301.20	295.91	281.49	1687.10
HEMOPEX	295.91	38.28	36.01	216.83
LACTDEHY	281.49	36.01	34.09	206.37
PYRUKIN	1687.10	216.83	206.37	1237.35

$varWi

$varWi[[1]]

	CREATKIN	HEMOPEX	LACTDEHY	PYRUKIN
CREATKIN	187.38	-17.66	2.20	-2.54
HEMOPEX	-17.66	55.44	3.67	54.66
LACTDEHY	2.20	3.67	8.77	27.44
PYRUKIN	-2.54	54.66	27.44	603.33

$varWi[[2]]

	CREATKIN	HEMOPEX	LACTDEHY	PYRUKIN
CREATKIN	8433	-6	879	1617
HEMOPEX	-6	41	-5	-19
LACTDEHY	879	-5	141	253
PYRUKIN	1617	-19	253	1515

b. $correlation

	CREATKIN	HEMOPEX	LACTDEHY	PYRUKIN
CREATKIN	1.00	0.22	0.82	0.55
HEMOPEX	0.22	1.00	0.22	0.38
LACTDEHY	0.82	0.22	1.00	0.62
PYRUKIN	0.55	0.38	0.62	1.00
	0.65	0.75	0.61	0.86

c. $statvar

	m CREATKIN	m HEMOPEX	m LACTDEHY	m PYRUKIN
1	4.31	7.97	1.26	16.50
2	15.56	9.42	2.64	24.75
moyenne	9.55	8.64	1.90	20.34
sd	12.32	1.37	1.60	6.83
F	18.96	28.00	16.35	41.46
P	0.00	0.00	0.00	0.00

d. $valpro

	[,1]
vp	9.955148e-01
F*	7.068155e+01
Bartlett	4.767224e+01
ddl	4.000000e+00
P	1.104557e-09

e. $fonction discriminante

```
              [,1]
[1,]  0.004352877
[2,]  0.059999982
[3,] -0.011854650
[4,]  0.013873564
```

f. $F

	[,1]
1	-0.062101255
2	-0.084733896
3	-0.063855878
4	0.109014047
5	-0.037833230
6	-0.137787748
7	0.013928659
8	-0.010936313
9	-0.121330774
10	-0.010704120
11	-0.094265690
12	-0.067923433
13	-0.317450539
14	-0.343866371
15	-0.288180337
16	-0.062324886
17	-0.023538650
18	0.098530693
19	-0.100461891
20	-0.067516810
21	-0.195069922
22	-0.131231614
23	-0.197560272
24	-0.109034745
25	-0.122281437
26	-0.105850763
27	-0.136813863
28	-0.028787680
29	-0.305945899
30	-0.212413567
31	-0.131256629
32	-0.207606220
33	-0.067626125
34	-0.092631049
35	-0.102081292
36	0.006698862
37	-0.168669399
38	-0.182875943
39	-0.087550185
40	0.260997859
41	0.019293866
42	0.224527797
43	0.041754192
44	-0.023331723
45	0.321524616
46	0.079661048
47	0.033765907
48	0.088409424
49	0.106028772
50	-0.030806550
51	0.081411763
52	-0.031118944
53	0.161602167
54	-0.115997061
55	-0.120691264
56	-0.023011731
57	0.200597201
58	0.456385625
59	0.400259046
60	0.338674929
61	0.086091840
62	0.287399204
63	0.128251977
64	0.158633650
65	0.223148810
66	0.155025059
67	0.053281349
68	0.180039195
69	0.088461811
70	0.034569732
71	0.197233673
72	0.089869013
73	0.099983908

g. $classe

ind	sc1	sc2	Ca	P	Err
1	0.06	-1.11	1	1	0
2	0.23	-1.31	1	1	0
3	0.07	-1.12	1	1	0
4	-1.27	0.41	1	2	1
5	-0.13	-0.89	1	1	0
6	0.64	-1.78	1	1	0
7	-0.53	-0.43	1	2	1
8	-0.34	-0.65	1	1	0
9	0.52	-1.63	1	1	0
10	-0.34	-0.65	1	1	0
11	0.31	-1.39	1	1	0
12	0.10	-1.16	1	1	0
13	2.03	-3.37	1	1	0
14	2.24	-3.61	1	1	0
15	1.81	-3.11	1	1	0
16	0.06	-1.11	1	1	0
17	-0.24	-0.77	1	1	0
18	-1.18	0.32	1	2	1
19	0.36	-1.45	1	1	0
20	0.10	-1.16	1	1	0
21	1.09	-2.29	1	1	0
22	0.59	-1.72	1	1	0
23	1.11	-2.31	1	1	0
24	0.42	-1.52	1	1	0
25	0.52	-1.64	1	1	0
26	0.40	-1.49	1	1	0
27	0.64	-1.77	1	1	0
28	-0.20	-0.81	1	1	0
29	1.95	-3.27	1	1	0
30	1.22	-2.44	1	1	0
31	0.59	-1.72	1	1	0
32	1.18	-2.40	1	1	0
33	0.10	-1.16	1	1	0
34	0.29	-1.38	1	1	0
35	0.37	-1.46	1	1	0
36	-0.47	-0.50	1	1	0
37	0.88	-2.05	1	1	0
38	0.99	-2.18	1	1	0
39	0.26	-1.33	1	1	0
40	-2.44	1.76	2	2	0
41	-0.57	-0.38	2	2	0
42	-2.16	1.44	2	2	0
43	-0.74	-0.19	2	2	0
44	-0.24	-0.76	2	1	1
45	-2.91	2.30	2	2	0
46	-1.04	0.15	2	2	0
47	-0.68	-0.26	2	2	0
48	-1.11	0.23	2	2	0
49	-1.24	0.39	2	2	0
50	-0.18	-0.83	2	1	1
51	-1.05	0.17	2	2	0
52	-0.18	-0.83	2	1	1
53	-1.67	0.88	2	2	0
54	0.48	-1.58	2	1	1
55	0.51	-1.63	2	1	1
56	-0.24	-0.76	2	1	1
57	-1.97	1.23	2	2	0
58	-3.96	3.50	2	2	0
59	-3.52	3.00	2	2	0
60	-3.05	2.45	2	2	0
61	-1.09	0.21	2	2	0
62	-2.65	2.00	2	2	0
63	-1.41	0.58	2	2	0
64	-1.65	0.85	2	2	0
65	-2.15	1.43	2	2	0
66	-1.62	0.82	2	2	0
67	-0.83	-0.08	2	2	0
68	-1.82	1.04	2	2	0
69	-1.11	0.23	2	2	0
70	-0.69	-0.25	2	2	0
71	-1.95	1.19	2	2	0
72	-1.12	0.24	2	2	0
73	-1.19	0.33	2	2	0

ind : n° individu Sc1 Sc2 : à définir : Ca : groupe d'origine P : ? Er: ?

h. $bilan

```
$bilan[[1]]
   gp
fac  1   2
  1  36  3
  2  6   28

$bilan[[2]]
[1] "Le % de mal classés est : 12.3287671232877"
```