ANNALES MA 412

EPMK

Année 2009-2010

MA412

Traitement statistique des données

Sujet

- 2006
- 2007
- 2008

Avertissement :

- ✓ Les programmes sont susceptibles d'être modifiés d'une année sur l'autre.
- ✓ Les corrections et les formulaires, aussi bien ceux des élèves comme ceux des professeurs peuvent contenir des erreurs.
- ✓ Pour plus de précision ou en cas de doutes, consulter le professeur responsable de l'enseignement.
- ✓ N'hésitez pas à consulter le site du BDE qui peut contenir des sujets et corrections supplémentaires.
- ✓ Pour toutes remarques ou suggestions, merci de me contacter par mail : thalyp@esiee.fr.

Bon courage

Thalyp

Groupe ESIEE **26/10/06**

O.de Cambry **MA-412**

Traitement statistique des données

Examen

Durée :**2 heures** Sujet à traiter **avec** documents

On considère le tableau de données présenté en annexe concernant des conditions climatiques à Bordeaux au cours des mois d'Avril à Septembre de 1924 à 1955 .

- A - Analyse en Composantes Principales

On a effectué une Analyse en Composantes Principales sur les données normalisées, dont les résultats sont rassemblés en Annexe.

1. *Calculer la fidélité de la représentation des données sur le plan principal.*
2. *Déterminer les corrélations entre les caractères initiaux et les 2 axes principaux et représenter les dans le cercle des corrélations (Figure 1).*
3. *Commenter les résultats de L'ACP.*

- B - Classification

1. On considère $\mathcal{C} = (M_1, \ldots, M_n)$ un nuage de points avec $x_i = (x_{i1}, \ldots, x_{ip})$ les coordonnées de M_i dans $\mathcal{R}^p$, $\mathcal{P} = (A_1, \ldots, A_k)$ une partition en k classes de $\mathcal{C}$. n_l le cardinal de A_l et V_l la covariance intra-classe (les définitions utiles sont rappelées en Annexe).
   Montrer que W s'exprime simplement en fonction de $(n_l)_{l=1,\ldots,k}$, les effectifs des classes et $(V_l)_{l=1,\ldots,k}$, les covariances intra-classes.
2. Dans la suite on travaille sur les données précédentes projetées dans le plan principal. On a classé les années en fonction de la qualité du vin obtenue en 3 classes :
   - **1** : Bonne année pour le vin
   - **2** : Année moyenne pour le vin
   - **3** : Année médiocre pour le vin

   Les informations nécessaires se trouvent en Annexe.

   (a) *Représenter les 3 classes sur le graphe de la projection des données sur le plan principal (Figure 2).*
   (b) *A l'aide de cette observation expliquer comment le climat influence la qualité du vin.*
   (c) *Calculer W pour cette classification.*
   (d) *Montrer que la classification proposée n'est pas optimale au sens du critère W.*

- C - Discrimination

On souhaite discriminer les 3 classes A_1, A_2 et A_3 en utilisant la mesure de voisinage de Mahalanobis.

1. *Déterminer le taux d'erreurs de la méthode.*
2. Les caractéristiques del'année 1956 est la suivante :

Année	X_1	X_2	X_3	X_4
56	3083	1195	5	441

   (a) *Déterminer les coordonnées du point représentant l' année 56 dans le plan principal et le placer sur la Figure 2.*

   (b) *A quelle classe doit-on l'affecter ? .*

Barème indicatif :

- A - : 6pts

- B - : 7pts

- C - : 7pts

Annexe

Données

Annees	X_1	X_2	X_3	X_4
24	3064	1201	10	361
25	3000	1053	11	338
26	3155	1133	19	393
27	3085	970	4	467
28	3245	1258	36	294
29	3267	1386	35	225
30	3080	966	13	417
31	2974	1185	12	488
32	3038	1103	14	677
33	3318	1310	29	427
34	3317	1362	25	326
35	3182	1171	28	326
36	2998	1102	9	349
37	3221	1424	21	382
38	3019	1239	16	275
39	3022	1285	9	303
40	3094	1329	11	339
41	3009	1210	15	536
42	3227	1331	21	414
43	3308	1368	24	282
44	3212	1289	17	302
45	3381	1444	25	253
46	3061	1175	12	261
47	3478	1317	42	259
48	3126	1248	11	315
49	3458	1508	43	286
50	3252	1361	26	346
51	3052	1186	14	443
52	3270	1399	24	306
53	3198	1299	20	367
54	2904	1164	6	311
55	3247	1277	19	375
Moyenne	3164.4	1251.7	19.4	357.6
Ecart-type	144.37	130.44	9.98	93.12

Variables

Les mesures ont été effectuées du 1^{er} Avril au 30 Septembre de chaque année.

- X_1 : somme des temperatures moyennes
- X_2 : ensoleillement en heures
- X_3 : nombre de jours de grande chaleur
- X_4 : hauteur de pluies en mm

A . C . P

Matrice de corrélation

$$R = \begin{pmatrix} 1.0000 & 0.7130 & 0.8686 & -0.4049 \\ 0.7130 & 1.0000 & 0.6475 & -0.4699 \\ 0.8686 & 0.6475 & 1.0000 & -0.3785 \\ -0.4049 & -0.4699 & -0.3785 & 1.0000 \end{pmatrix}$$

Valeurs propres

$$\begin{aligned} \lambda_1 &= 2.78 \\ \lambda_2 &= 0.73 \\ \lambda_3 &= 0.36 \\ \lambda_4 &= 0.13 \end{aligned}$$

Vecteurs propres

u_1	u_2	u_3	u_4
-0.553	0.292	0.215	0.750
-0.515	-0.005	-0.846	-0.136
-0.537	0.332	0.428	-0.647
0.376	0.897	-0.233	-0.006

Classes

Notations

Soit $\mathcal{C} = (M_1, \ldots, M_n)$ avec $x_i = (x_{i1}, \ldots, x_{ip})$ les coordonnées de M_i dans $\mathcal{R}^p$, $\mathcal{P} = (A_1, \ldots, A_k)$ une partition en k classes de $\mathcal{C}$ et n_l le cardinal de A_l.

- $G_l = (\bar{x}_{l1}, \ldots, \bar{x}_{lp}) = (\frac{1}{n_1} \sum_{\{i, M_i \in A_1\}} x_{i1}, \ldots, \frac{1}{n_k} \sum_{\{i, M_i \in A_k\}} x_{ip})$ centre de gravité de la classe A_l
- $I_l = I(A_l) = \frac{1}{n_l} \sum_{\{i, M_i \in A_l\}} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{lj})^2$ inertie de la classe A_l
- $V_l = [\gamma_{jj'}]$ avec $\gamma_{jj'} = \frac{1}{n_l} \sum_{\{i, M_i \in A_l\}} (x_{ij} - \bar{x}_{lj})(x_{ij'} - \bar{x}_{lj'})$ matrice de covariance de la classe A_l
- $W(\mathcal{P}) = \sum_{l=1}^{k} n_l I_l = \sum_{l=1}^{k} \sum_{\{i, M_i \in A_l\}} d^2(G_l, M_i)$ critère de la somme des inerties

Classes et distance au centre des classes

Annees	Classe	$d(M, G_1)$	$d(M, G_2)$	$d(M, G_3)$
24	2	8.51	0.83	1.66
25	2	13.43	2.76	1.56
26	2	5.91	0.63	1.83
27	3	20.06	6.81	0.49
28	1	0.05	3.19	12.59
29	1	0.78	6.59	20.00
30	3	14.51	3.69	0.05
31	3	13.89	3.92	0.05
32	3	26.42	15.24	5.56
33	2	1.96	4.41	10.39
34	1	0.08	3.15	12.30
35	2	1.95	0.45	5.79
36	3	13.13	2.62	1.50
37	1	0.81	1.80	8.56
38	2	6.56	0.84	5.16
39	2	7.87	1.00	4.17
40	2	4.84	0.09	3.91
41	3	13.56	4.92	0.81
42	2	2.16	1.55	6.07
43	1	0.17	3.28	13.58
44	2	2.09	0.47	6.96
45	1	0.60	6.62	19.83
46	2	8.26	1.39	5.07
47	1	2.04	11.79	26.50
48	2	5.53	0.25	3.95
49	1	4.42	16.87	33.48
50	2	0.34	2.38	10.35
51	3	9.63	1.84	0.46
52	1	0.13	2.89	12.51
53	1	2.04	0.55	5.54
54	3	15.62	4.05	3.50
55	1	2.04	0.74	5.56

Centre des classes

- G_1=(-1.7 ; 0.1)
- G_2=(0.2 ; -0.3)
- G_3=(2 ; 0.4)

Covariance intraclasse

$$V1 = \begin{pmatrix} 1 & 0 \\ 0 & 0.1 \end{pmatrix} \qquad V2 = \begin{pmatrix} 0.7 & -0.4 \\ -0.4 & 0.6 \end{pmatrix} \qquad V3 = \begin{pmatrix} 0.2 & 0.2 \\ 0.2 & 1.4 \end{pmatrix}$$

Mesure de voisinage de Mahalanobis

Rappel : si G_A est le centre de gravité de la classe A et si V_A est sa matrice de covariance, la mesure de voisinage de Mahalanobis entre un point M et la classe A est définie par $S_A(M) = (M - G_A)^t \, V_A^{-1} \, (M - G_A)$.

Annees	Classe	$S_1(M)$	$S_2(M)$	$S_3(M)$
24	2	9.43	1.66	4.20
25	2	16.34	4.48	1.12
26	2	6.24	2.90	11.01
27	3	20.34	24.74	3.11
28	1	0.06	5.55	75.55
29	1	3.60	19.39	110.67
30	3	14.12	13.39	0.07
31	3	16.14	16.74	0.11
32	3	73.66	70.50	4.34
33	2	11.89	4.47	65.99
34	1	0.16	5.00	74.61
35	2	1.84	0.46	33.22
36	3	15.74	4.36	1.08
37	1	1.53	1.81	53.16
38	2	15.73	1.67	12.51
39	2	16.16	1.17	8.19
40	2	7.02	0.14	15.07
41	3	23.93	22.79	1.63
42	2	5.16	1.88	38.80
43	1	1.04	8.27	76.58
44	2	4.00	1.70	33.77
45	1	2.19	17.86	112.66
46	2	20.23	2.05	8.89
47	1	2.51	21.31	165.16
48	2	9.13	0.34	13.02
49	1	5.96	29.20	211.17
50	2	0.63	3.08	63.45
51	3	10.41	8.14	2.88
52	1	0.41	6.31	72.27
53	1	2.06	0.51	33.04
54	3	27.03	4.08	2.77
55	1	2.47	0.73	34.11

Ne pas oublier de rendre cette page avec la copie

Figure 1

NOM :

PRENOM :

Question - A - 2

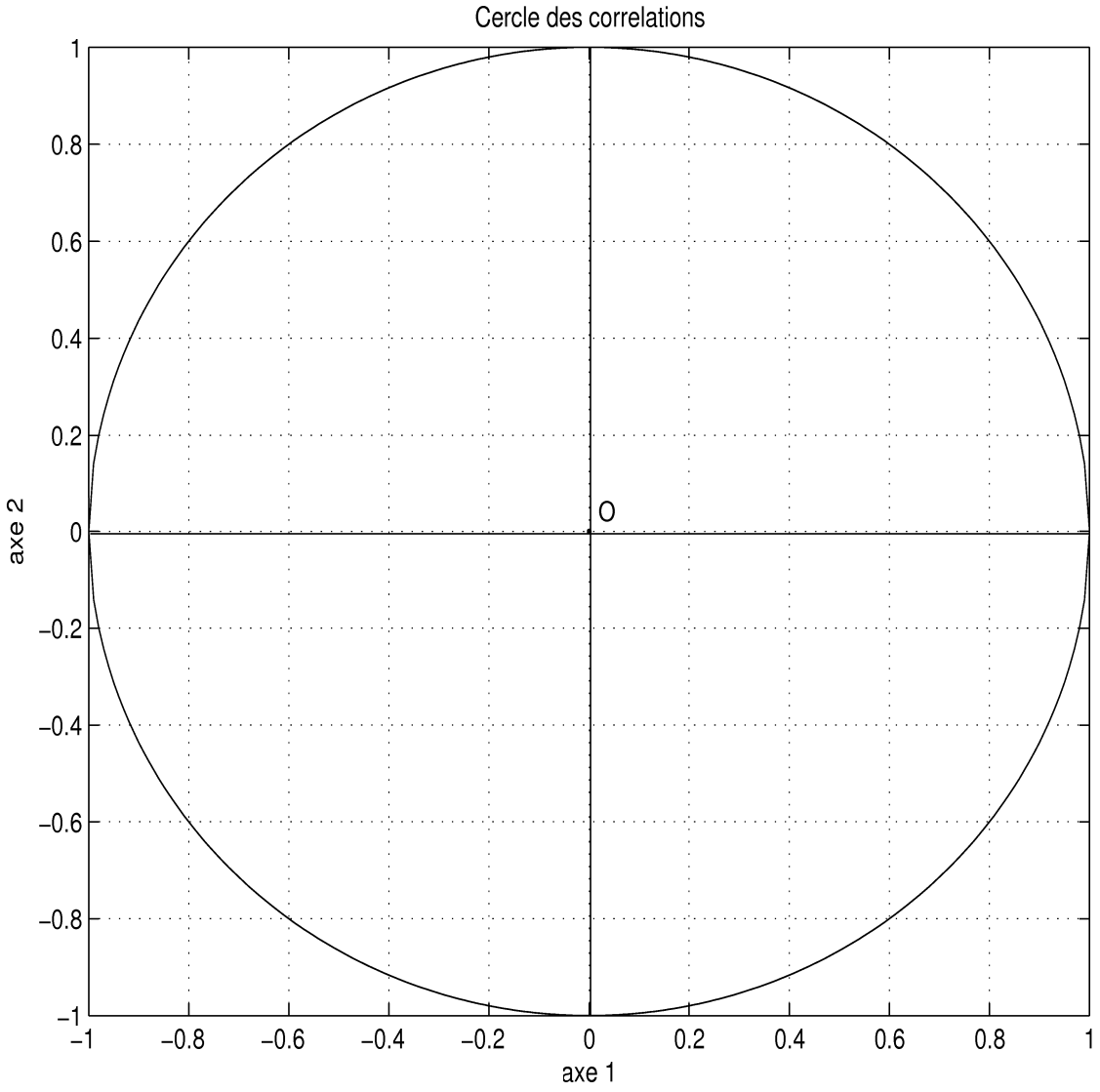


FIG. 1 – Cercle de corrélation

Question - B - 1

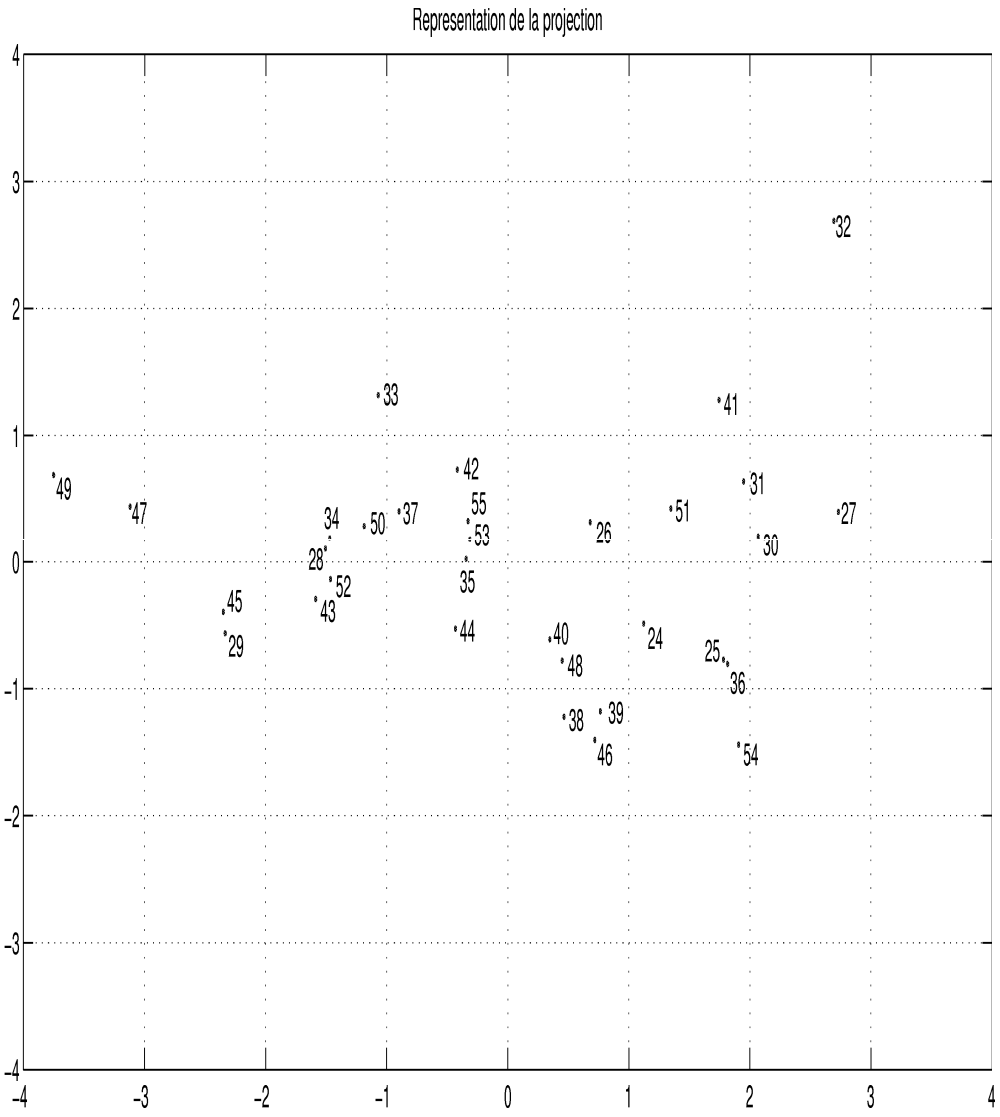


FIG. 2 – Projection sur le plan principal

Mathématiques

- A - Analyse en Composantes Principales -

1 - Fidélité : $F = \frac{1}{p}\sum_{k=1}^{p} \lambda_k = \frac{1}{4}(\lambda_1 + \lambda_2) = \frac{1}{4}(2,78 + 0,73)$

$= 0,8775$ soit $87,75\%$

La fidelité de la représentation dans le plan principal (u_1, u_2) est de 87,75%. Avec les 2 axes principaux on a donc une vision quasi complète des données.

2 - Déterminons les corrélations entre les caractères initiaux et les 2 axes principaux.

x_1 $\rho(X_1, u_1(X_1)) = \sqrt{\lambda_1}\, u_1^1 = \sqrt{2,78} \times -0,553 = -0,922$

y_1 $\rho(X_1, u_2(X_1)) = \sqrt{\lambda_2}\, u_2^1 = \sqrt{0,73} \times 0,292 = 0,2495$

x_2 $\rho(X_2, u_1(X_2)) = \sqrt{\lambda_1}\, u_1^2 = \sqrt{2,78} \times -0,515 = -0,8587$

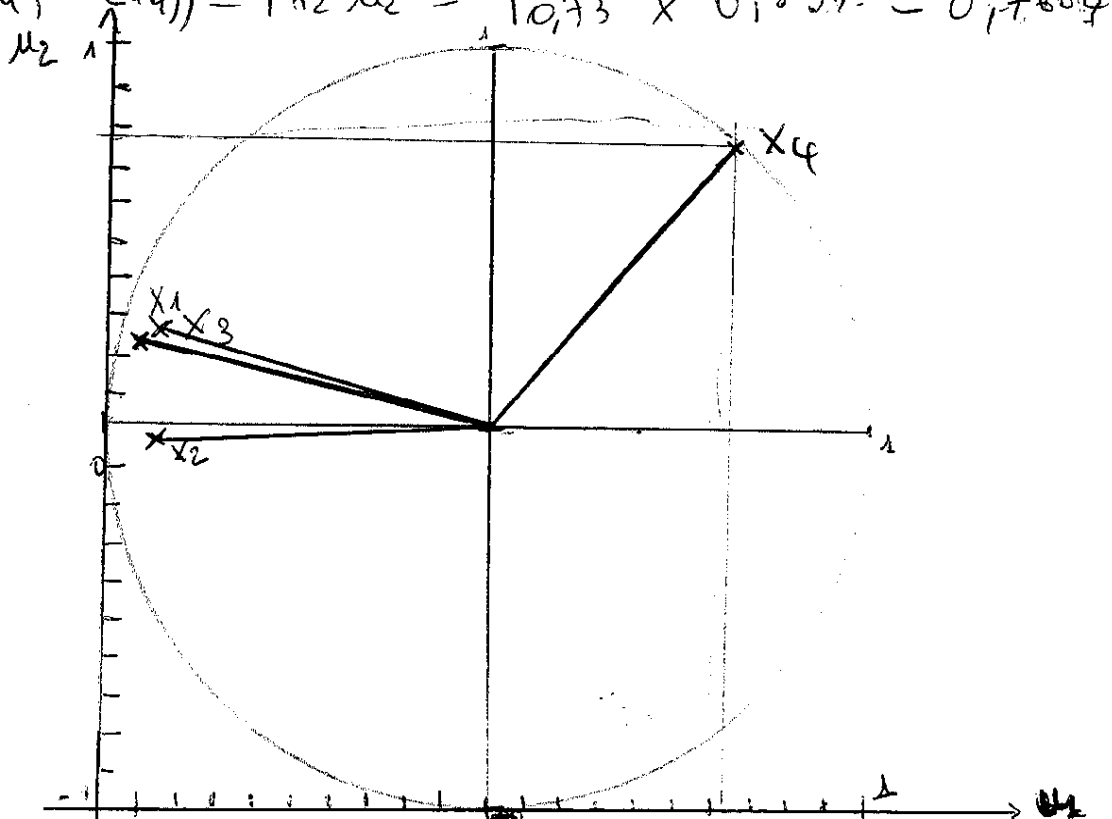
y_2 $\rho(X_2, u_2(X_2)) = \sqrt{\lambda_2}\, u_2^2 = \sqrt{0,73} \times -0,005 = -0,00427$

x_3 $\rho(X_3, u_1(X_3)) = \sqrt{\lambda_1}\, u_1^3 = \sqrt{2,78} \times -0,537 = -0,89536$

y_3 $\rho(X_3, u_2(X_3)) = \sqrt{\lambda_2}\, u_2^3 = \sqrt{0,73} \times 0,332 = 0,28366$

x_4 $\rho(X_4, u_1(X_4)) = \sqrt{\lambda_1}\, u_1^4 = \sqrt{2,78} \times 0,376 = 0,6269$

y_4 $\rho(X_4, u_2(X_4)) = \sqrt{\lambda_2}\, u_2^4 = \sqrt{0,73} \times 0,897 = 0,7664$



Un vecteur de grande longueur proche d'un axe indique son poids dans la définition de l'axe : on peut considérer que c'est le cas de X_2 : l'ensoleillement en heure pour l'axe 1.

On remarque que les vecteurs correspondants à la somme des températures moyennes (X_1) et le nombre de jours de grande chaleur (X_3) sont longs et très proches l'un de l'autre. On peut en déduire qu'elles sont fortement positivement corrélées.

En revanche il semblerait qu'il n'y ait pas de corrélation entre $X_1, X_2 X_3$ et X_4 : la hauteur de pluie en mm, car les variables sont orthogonales quasiment les unes par rapport à l'autre.

B - Classification

1. $C = (M_1, \dots, M_N)$, $x_i = (x_{i1}, \dots x_{ip})$; $P = (A_1, \dots, A_k)$.

n_ℓ cardinal de A_ℓ et V_ℓ la covariance intraclasse.

Mentions que W s'exprime simplement en fonction de $(n_\ell)_{\ell=1,\dots,k}$ les effectifs des classes et $(V_\ell)_{\ell=1\dots k}$, les covariances intra-classes.

$$W(P) = \sum_{\ell=1}^{k} n_\ell I_\ell = \sum_{\ell=1}^{k} n_\ell \times \frac{1}{n_\ell} \sum_{\{i, M_i \in A_\ell\}} \sum_{j=1}^{p} (x_{ij} - \overline{x_{\ell j}})^2$$

$$= \sum_{\ell=1}^{k} n_\ell \times \sum_{j=1}^{p} \left[\frac{1}{n_\ell} \sum_{\{i, M_i \in A_\ell\}} (x_{ij} - \overline{x_{\ell j}})^2 \right]$$

$$= \sum_{\ell=1}^{k} n_\ell \sum_{j=1}^{p} \gamma_{jj} = \sum_{\ell=1}^{k} n_\ell \times T_r(V_\ell)$$

2- (a) avec des couleurs on distingue les 3 classes sur le graphe de la projection des données sur le plan principal.

(b) Le climat influence la qualité du vin : ....

(c) $$W = \sum_{l=1}^{k} n_l T_R(V_l) = n_1 T_R(V_1) + n_2 T_R(V_2) + n_3 T_R(V_3)$$

$$= 11 \times 1,1 + 13 \times 1,3 + 8 \times 1,6$$

$$= 41,8$$

(d) Montrons que la classification n'est pas optimale au sens du critère W : l'année 1925 est plus proche de la classe A3 que de la classe A2 où elle a été affectée, on peut donc minimiser W.

C - Discrimination

Discriminons les trois classes A_1, A_2 et A_3 en utilisant la mesure de voisinage de Mahalanobis.

1 - Déterminons le taux d'erreur de la méthode:

Pour cela déterminons le nombre de points mal classés par classe.

$$\Sigma_{ll'} = \{M \in \mathbb{R}^2, (M-G_l)^t V_l^{-1} (M-G_l) = (M-G_{l'}) V_{l'}^{-1} (M-G_{l'})\}$$

$$\begin{pmatrix} x+1,7 \\ y-0,1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 10 \end{pmatrix} \begin{pmatrix} x+1,7 & y-0,1 \end{pmatrix}$$

$\det V_1 = 0,1$

$\frac{1}{\det V_1} = 10$.

$$= (x+1,7)^2 + (10y-1)(y-0,1)$$

$$= x^2 + 3,4x + 2,89 + 10y^2 - y - y + 0,1$$

$$= x^2 + 10y^2 + 3,4x - 2y + 2,99$$

$$\frac{1}{0,26} \begin{pmatrix} x-0,2 & y+0,3 \end{pmatrix} \begin{pmatrix} 0,6 & 0,4 \\ 0,4 & 0,7 \end{pmatrix} \begin{pmatrix} x-0,2 \\ y+0,3 \end{pmatrix}$$

$$= \frac{1}{0,26} \begin{bmatrix} x-0,2 & y+0,3 \end{bmatrix} \begin{bmatrix} 0,6x+0,4y \\ 0,4x+0,7y+0,13 \end{bmatrix}$$

$$= \frac{1}{0,26} (x-0,2)(0,6x+0,4y) + (y+0,3)(0,4x+0,7y+0,13)$$

$$= \frac{1}{0,26} (0,6x^2 + 0,4xy - 0,12x - 0,8y + 0,4xy + 0,7y^2 + 0,3y + 0,12x + 0,21y + 0,039)$$

$$= \frac{1}{0,26} (0,6x^2 + 0,7y^2 + xy(0,8) + x(-0,12+0,12) + y(-0,8+0,3-0,46) + 0,039 \ldots$$

C - Discrimination.

1) Nb de points mal classés : 4

taux d'erreur : $\frac{1}{32}\left(2 \times \frac{11}{32} + 2 \times \frac{13}{32} + 0 \times \frac{8}{32}\right)$

$= \frac{1}{32}\left(\frac{22}{32} + \frac{26}{32}\right) = 0,046$

$= 4,68\ \%$

2] Année 1956.

a) en normalise

$$\begin{pmatrix} \frac{3083 - 3164,4}{144,37} \\ \frac{1195 - 1254,7}{130,44} \\ \frac{5 - 19,4}{9,98} \\ \frac{441 - 357,6}{93,12} \end{pmatrix} = \begin{pmatrix} -0,5638 \\ -0,4347 \\ -1,4429 \\ 0,8956 \end{pmatrix}$$

on projette.

$\pi_R(1956) = \langle 1956_N, u_R \rangle$

$\pi_1(1956) = -0,5638 \times -0,553 + -0,4347 \times -0,515 + -1,4429 \times -0,5$

$+ 0,8956 \times 0,376 = 1,647$

$\pi_2(1956) = -0,5638 \times 0,292 + -0,4347 \times -0,005 + -1,4429 \times 0,332 + 0,8956 \times 0,8$

$= 0,162$

2] A quelle classe doit-on affecter l'année 1956.

$$S_1(M_{1956}) = (M_{1956} - G_1)^t \, V_1^{-1} (M_{1956} - G_1)$$

$$\begin{bmatrix} 3,347 \\ 0,062 \end{bmatrix}$$

$$\begin{bmatrix} 1,647+1,7 & 0,162-0,1 \end{bmatrix} \times \frac{1}{0,1} \begin{bmatrix} 0,1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$10 \begin{bmatrix} 3,347 & 0,062 \end{bmatrix} \begin{bmatrix} 0,347 \\ 0,062 \end{bmatrix} = 10 \begin{bmatrix} 3,347 \times 0,3347 + 0,062 \times 0,062 \end{bmatrix}$$

$$= 11,24 \qquad \begin{bmatrix} 1,447 \\ 0,462 \end{bmatrix}$$

$$S_2(M_{1956}) = \begin{bmatrix} 1,647-0,2 & 0,162+0,3 \end{bmatrix} \times \frac{1}{0,26} \begin{bmatrix} 0,6 & 0,4 \\ 0,4 & 0,7 \end{bmatrix}$$

$$= \frac{1}{0,26} \begin{bmatrix} 1,447 & 0,462 \end{bmatrix} \begin{bmatrix} 0,6 \times 1,447 + 0,4 \times 0,462 \\ 0,4 \times 1,447 + 0,7 \times 0,462 \end{bmatrix}$$

$$= \frac{1}{0,26} \begin{bmatrix} 1,447 & 0,462 \end{bmatrix} \begin{bmatrix} 1,053 \\ 0,9022 \end{bmatrix} = \frac{1}{0,26} \begin{bmatrix} 1,447 \times 1,053 + \\ 0,462 \times 0,9022 \end{bmatrix}$$

$$= 7,46349$$

$$S_3(M_{1956}) = \begin{bmatrix} 1,647-2 & 0,162-0,4 \end{bmatrix} \times \frac{1}{0,24} \begin{bmatrix} 1,4 & -0,2 \\ -0,2 & 0,2 \end{bmatrix} \begin{bmatrix} -0,353 \\ -0,238 \end{bmatrix}$$

$$= \frac{1}{0,24} \begin{bmatrix} -0,353 & -0,238 \end{bmatrix} \begin{bmatrix} 1,4 \times -0,353 + -0,2 \times -0,238 \\ = -0,4466 \\ -0,2 \times -0,353 + 0,2 \times -0,238 \\ = 0,023 \end{bmatrix}$$

$$= 0,634.$$

On l'affecte à la classe A(3).

C - Discrimination (suite)

2. Les caractéristiques de l'année 1956 sont

$X_1 = 3083 \quad X_2 = 1195 \quad X_3 = 5 \quad X_4 = 441.$

Pour projeter dans le plan principal (si on ne normalise pas)

$\Pi_k(1956) = \langle 1956', u_k \rangle \Rightarrow \begin{pmatrix} 3083 \\ 1195 \\ 5 \\ 441 \end{pmatrix}$ u_1 $\begin{pmatrix} -0,553 \\ -0,515 \\ -0,537 \\ 0,376 \end{pmatrix}$ u_2 $\begin{pmatrix} 0,292 \\ -0,005 \\ 0,332 \\ 0,897 \end{pmatrix}$

$\Pi_1(9956) = -2157,193$

$\Pi_2(1956) = 1291,5$

Si on normalise il faut recalculer moyenne et écart-type.

---

1) taux d'erreur $= \frac{1}{n_t}\left(n_1 \times \frac{n_{e1}}{n_{t1}} + n_2 \frac{n_{e2}}{n_{t2}} + n_3 \times \frac{n_{e3}}{n_{t3}}\right)$

n_i : nb de pts mal classés de la classe i

taux d'erreur $= \frac{1}{32}\left(1 \times \frac{11}{32} + (2) \times \frac{13}{32} + 0 \times \frac{8}{32}\right)$

$= \frac{1}{32}\left(\frac{11}{32} + \frac{2 \times 13}{32}\right) = 3,61\,\%$

ou : nb de points mal classés (plus proches d'une autre classe : 4.

Groupe ESIEE **07/11/07**

O.de Cambry **MA-412**

Traitement statistique des données

Examen

Durée :**2 heures** Sujet à traiter **avec** documents

- A - Présentation des données

On considère le tableau de données suivant concernant 20 villes européennes.

Numéro	Ville	X_1	X_2	X_3	X_4	X_5
1	Amsterdam	52.38	4.92	1.5	17.5	765
2	Athènes	37.97	23.72	9.5	28	398
3	Barcelone	41.4	2.15	9.5	24.5	600
4	Berlin	52.45	13.2	-0.5	19	610
5	Copenhague	55.68	12.55	0	18	605
6	Dublin	53.37	-6.35	4.5	15.5	755
7	Helsinki	60.32	24.9	-6.1	16.8	635
8	Lisbonne	38.72	-9.13	11	22	681
9	Londres	51.48	0	4	18	595
10	Lyon	45.7	4.78	3	20.5	810
11	Madrid	40.45	-3.50	5.3	24.6	439
12	Milan	45.43	9.19	1.1	23.8	984
13	Oslo	59.93	10.73	-4.5	17.5	735
14	Paris	48.82	2.48	4	19.5	585
15	Prague	50.08	14.42	-2	18.5	525
16	Rome	41.8	12.6	8	25	740
17	Sofia	42.7	23.33	-1	21.5	665
18	Stockholm	59.35	18.07	-3	18	560
19	Varsovie	52.40	21.00	-3.1	18.8	446
20	Vienne	48.25	16.37	-1.5	20	654
Moyenne		48.93	9.77	1.99	20.4	639
Ecart-type		7.01	10.12	4.94	3.32	138

Les variables étudiées sont :

- X_1 :Latitude (en degré)
- X_2 :Longitude(en degré)
- X_3 :Temperature moyenne de Janvier (en degré Celsius)
- X_4 :Temperature moyenne de Juillet (en degré Celsius)
- X_5 :Hauteur annuelle des precipitations (en mm)

- A - Analyse des résultats

On a effectué une Analyse en Composantes Principales sur les données normalisées, dont les résultats sont rassemblés à la suite des questions.

1. *Calculer la fidélité de la représentation des données sur le plan principal.*
2. *Déterminer les corrélations entre les caractères X_1 et X_2 et les 2 axes principaux et représenter les dans le cercle des corrélations (Figure 1).*
3. *Commenter les résultats de L'ACP.*

- B - Classification

Dans la suite on travaille sur la projection des données normalisées dans le plan principal. Les coordonnées des points se trouvent en annexe.

1. Soient A et B, 2 groupements de points disjoints, on définit la distance du lien maximum entre ces 2 groupements par : $D(A,B) = \max\limits_{M\in A, M'\in B} d(M,M')$

   Montrer que si A, B et C sont trois parties disjointes de C,
   $D(A\cup B, C) = \max(D(A,C), D(B,C))$.

2. La hiérarchie définie à partir du critère d'aggrégation du lien maximum a été construite par l'algorithme de la construction ascendante hiérarchique jusqu'à l'étape 16. Les résultats obtenus sont présentés Figure 2.
   On note $\mathcal{P}_{16} = (A_1, A_2, A_3, A_4, A_5)$ la partition obtenue. Sa composition ainsi que le tableau des distances entre ces groupements de points est donnée en Annexe.

   (a) *Continuer l'algorithme à partir de cette étape : pour chaque étape on donnera le tableau des distances entre classes et l'indice d'aggrégation et on complétera le dendrogramme (Figure 2).*

   (b) *Déterminer la classification en 2 classes associée.*

3. On considère à présent la partition en 2 classes d'un ensemble $\mathcal{C}$ de n points de $\mathbb{R}^p$ composée d'une classe C_1 réduite à un point noté M_0 et d'une autre classe C_2 composée de tous les autres points : $\mathcal{P} = \{\{M_0\}, \{\mathcal{C}\setminus\{M_0\}\}$. On note $I_\mathcal{C}$ l'inertie du nuage et $W(\mathcal{P})$ la valeur du critère de la somme des inerties pour la partition $\mathcal{P}$.

   (a) *Montrer que $W(\mathcal{P}) = n\, I_\mathcal{C} - \frac{n-2}{n-1}\, ||GM_0||^2$.*

   (b) On note à présent $\mathcal{P}_1$ la partition composée de la ville d'Athènes toute seule pour la classe C_1 et des autres villes pour la classe C_2.
   Calculer le critère de la somme des inerties W_1, pour cette partition.

- C - Discrimination

On considère maintenant la partition en 2 classes des données dans le plan principal définie ci-dessous :

- C_1={Amsterdam, Berlin, Copenhague, Dublin, Helsinki, Londres, Lyon, Milan, Oslo, Paris, Prague, Sofia, Stockholm, Varsovie, Vienne }
- C_2={Athènes, Barcelone, Lisbonne, Madrid, Rome }

Tous les éléments numériques utiles sont en Annexe.

1. *Déterminer l'inertie des classes C_1 et C_2 et en déduire la valeur du critère W_2 de la somme des inerties pour cette partition. Comparer avec W_1.*
   (Indication : Réfléchissez avant de vous lancer le calcul de W_2 est très simple)
2. On souhaite discriminer les 2 classes C_1 et C_2 en utilisant la distance entre centres pondérée par l'inertie.
   Soit $\alpha = \dfrac{n_2\, I_2}{n_1\, I_1}$ et Ω, le point défini par $\overrightarrow{\Omega G_2} = \alpha\, \overrightarrow{\Omega G_1}$ où G_1 et G_2 sont les centres de gravité respectifs de C_1 et C_2.
   On rappelle que la courbe séparatrice Σ_{12} est le cercle de centre Ω et de rayon $r = \dfrac{\sqrt{\alpha}}{|1-\alpha|}\, d(G_1, G_2)$.
   (a) *Déterminer a, Ω et r.*
   (b) *La ville d' Athènes est-elle correctement classée par cette mesure de voisinage ?*
   (c) La ville de Toulouse a pour caractéristiques : $M_{Toulouse}$ = [43.6 1.43 4.7 20.9 656]. *Déterminer la position de Toulouse sur le plan principal. A quelle classe doit-on l'affecter ?*

Barème indicatif :

- A - : 6pts

- B - : 7pts

- C - : 7pts

Ne pas oublier de rendre les pages 7 et 8 avec la copie

Annexe

- A -

Matrice de corrélation

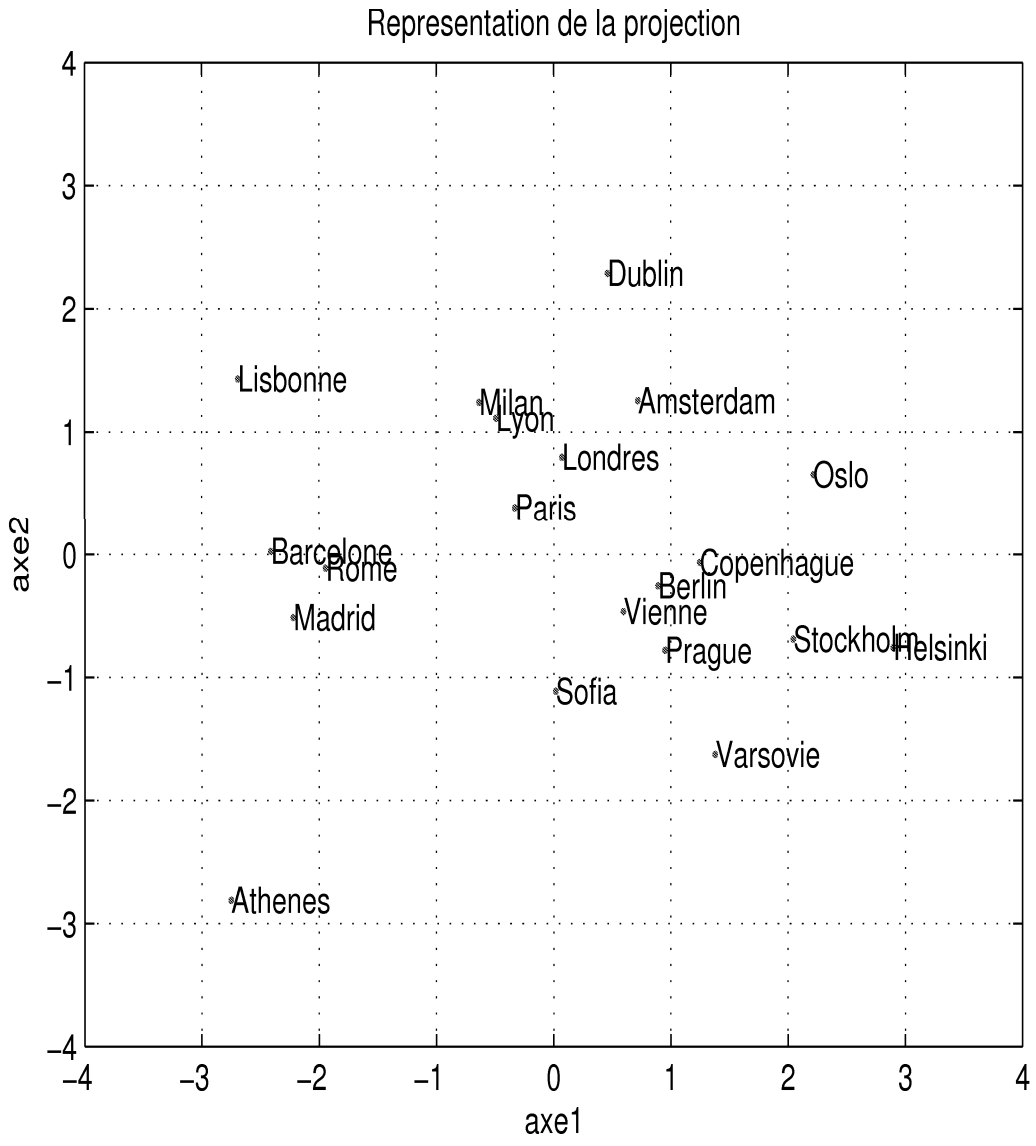
$$R = \begin{pmatrix} 1.0000 & 0.2605 & -0.8004 & -0.8653 & 0.0787 \\ 0.2605 & 1.0000 & -0.5796 & 0.0302 & -0.2521 \\ -0.8004 & -0.5796 & 1.0000 & 0.6605 & -0.0487 \\ -0.8653 & 0.0302 & 0.6605 & 1.0000 & -0.1486 \\ 0.0787 & -0.2521 & -0.0487 & -0.1486 & 1.0000 \end{pmatrix}$$

Valeurs propres

$$\begin{aligned} \lambda_1 &= 2.69 \\ \lambda_2 &= 1.34 \\ \lambda_3 &= 0.77 \\ \lambda_4 &= 0.12 \\ \lambda_4 &= 0.08 \end{aligned}$$

Vecteurs propres

$$U = \begin{pmatrix} 0.5780 & 0.1116 & 0.1664 & 0.5727 & 0.5457 \\ 0.2676 & -0.6700 & -0.4851 & 0.3416 & -0.3570 \\ -0.5691 & 0.1509 & 0.1921 & 0.7397 & -0.2631 \\ -0.5177 & -0.3273 & -0.3480 & 0.0112 & 0.7097 \\ 0.0489 & 0.6393 & -0.7609 & 0.0891 & -0.0440 \end{pmatrix}$$


Representation de la projection
Dublin
Lisbonne
Milan
Amsterdam
Lyon
Londres
Oslo
Paris
Barcelone
Rome
Copenhague
Berlin
Madrid
Vienne
Stockholm
Helsinki
Prague
Sofia
Varsovie
Athenes
axe1
axe2

- B -

Distance entre les classes à l'étape 16

- A_1={Lyon, Milan, Londres, Paris, Amsterdam, Dublin }
- A_2={Barcelone, Rome, Lisbonne, Madrid}
- A_3={Berlin, Vienne, Prague, Copenhague, Sofia, Varsovie}
- A_4={Helsinki, Stockholm, Oslo }
- A_5={Athènes }

A_1	0				
A_2	3.90	0			
A_3	4.02	5.09	0		
A_4	4.06	6.00	2.90	0	
A_5	6.03	4.24	4.85	6.05	0
	A_1	A_2	A_3	A_4	A_5

- C -

Coordonnées dans le plan principal et caractéristiques des classes

C_1	Nom	Xp_1	Xp_2	C_2	Nom	Xp_1	Xp_2
1	Amsterdam	0.72	1.26	2	Athènes	-2.75	-2.81
4	Berlin	0.89	-0.26	3	Barcelone	-2.41	0.02
5	Copenhague	1.24	-0.07	8	Lisbonne	-2.69	1.43
6	Dublin	0.46	2.29	11	Madrid	-2.22	-0.51
7	Helsinki	2.90	-0.76	16	Rome	-1.95	-0.11
9	Londres	0.07	0.80				
10	Lyon	-0.49	1.11				
12	Milan	-0.63	1.24				
13	Oslo	2.21	0.65				
14	Paris	-0.33	0.38				
15	Prague	0.95	-0.78				
17	Sofia	0.02	-1.11				
18	Stockholm	2.04	-0.69				
19	Varsovie	1.38	-1.63				
20	Vienne	0.59	-0.46				
Moyenne		0.80	0.13	Moyenne		-2.4	-0.39
Ecart-type		0.99	1.04	Ecart-type		0.3	1.37

Ne pas oublier de rendre cette page avec la copie

Figure 1

NOM :

PRENOM :

Question - A - 2

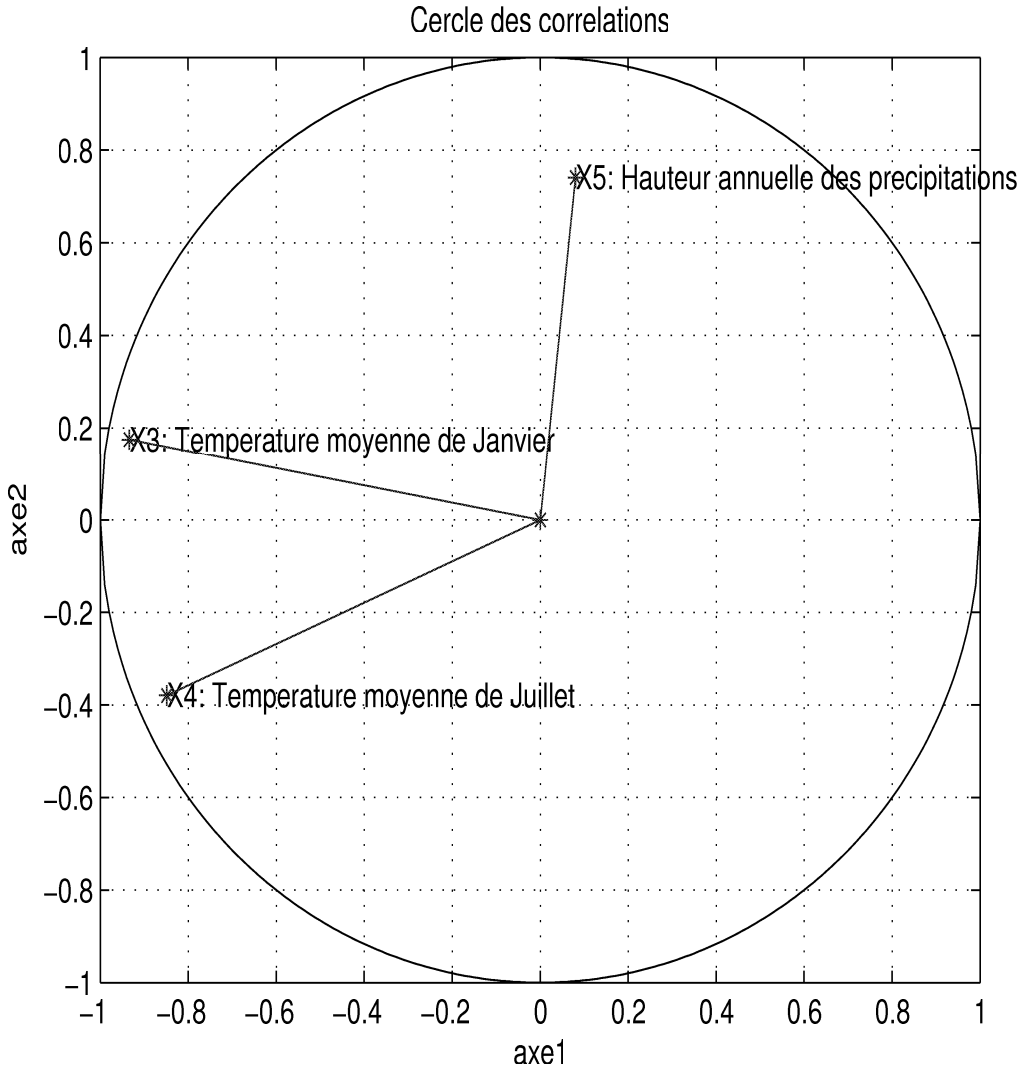


FIG. 1 – Cercle de corrélation

Ne pas oublier de rendre cette page avec la copie

Figure 1

NOM :

PRENOM :

Question - B - 2

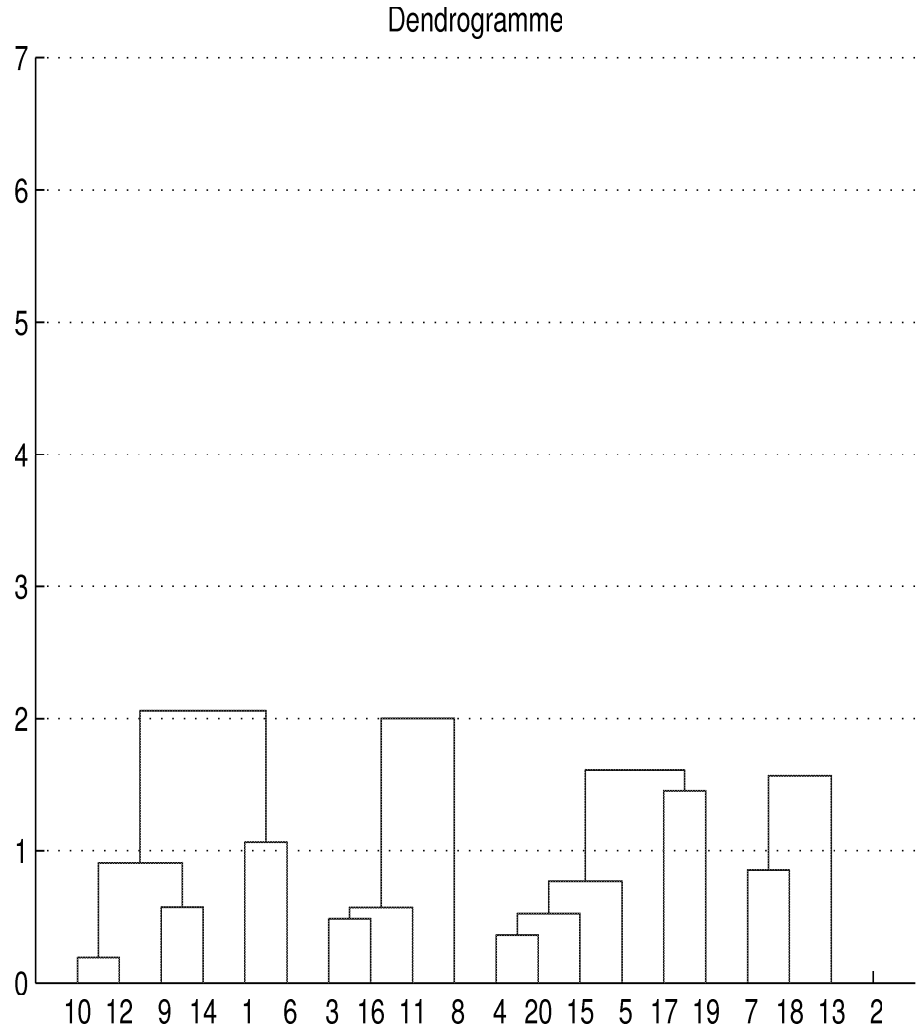


FIG. 2 – Projection sur l'axe principal

A - Analyse des résultats

1. Calculons la fidélité sur le plan principal : $F = \frac{1}{p}\sum_{k=1}^{q} d_k = \frac{1}{5}(\lambda_1 + d_2)$

$= \frac{1}{5}(2,69 + 1,34)$

2. Détermination des corrélations :

$\rho(X_1, u_1(X_1)) = \sqrt{\lambda_1}\, u_1^1 = \sqrt{2,69} \times 0,5780 = 0,948$

$\rho(X_1, u_2(X_1)) = \sqrt{\lambda_2}\, u_2^1 = \sqrt{1,34} \times 0,1116 = 0,134$

$\rho(X_2, u_1(X_2)) = \sqrt{\lambda_1}\, u_1^2 = \sqrt{2,69} \times 0,2676 = 0,439$

$\rho(X_2, u_2(X_2)) = \sqrt{\lambda_2}\, u_2^2 = \sqrt{1,34} \times -0,6700 = -0,775$

$\rho(X_3, u_1(X_3)) = \sqrt{\lambda_1}\, u_1^3 = \sqrt{2,69} \times -0,5691 = -0,933$

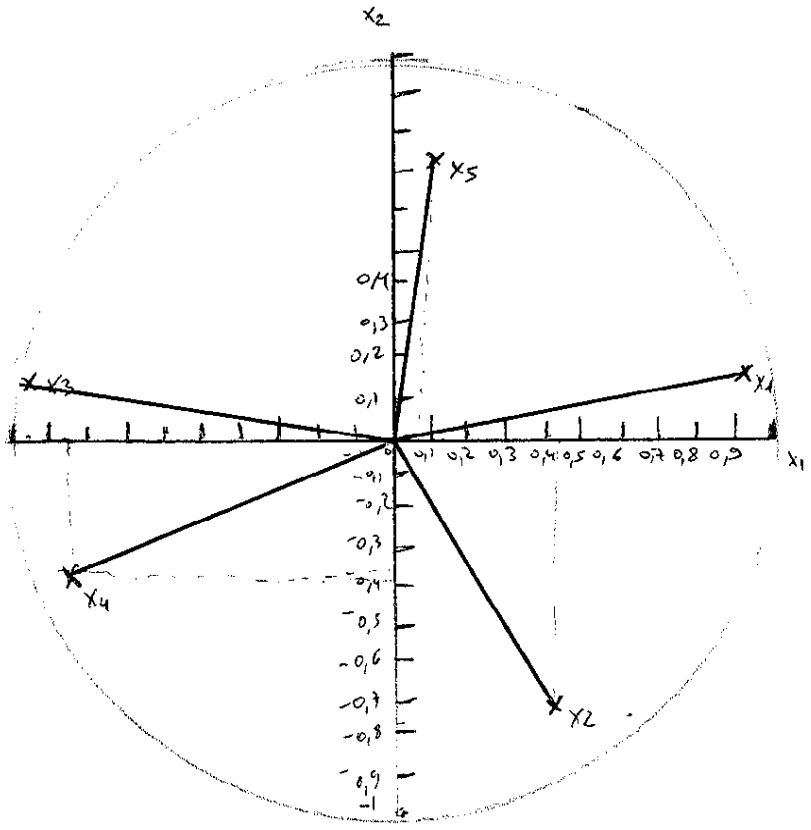
$\rho(X_3, u_2(X_3)) = \sqrt{\lambda_2}\, u_2^3 = \sqrt{1,34} \times 0,1509 = 0,175$

$\rho(X_4, u_1(X_4)) = \sqrt{\lambda_1}\, u_1^4 = \sqrt{2,690} \times -0,5177 = -0,849$

$\rho(X_4, u_2(X_4)) = \sqrt{\lambda_2}\, u_2^4 = \sqrt{1,34} \times -0,3273 = -0,379$

$\rho(X_5, u_1(X_5)) = \sqrt{\lambda_5}\, u_1^5 = \sqrt{2,69} \times 0,0489 = 0,080$

$\rho(X_5, u_2(X_5)) = \sqrt{\lambda_5}\, u_2^5 = \sqrt{1,34} \times 0,6393 = 0,740$



3) La fidélité de représentation dans le plan principal (u_1, u_2) est de 0,806. Soit 80,6 %, avec les deux axes principaux on a donc une vision quasi complète des données.

Un vecteur de grande longueur proche de l'axe indique son poids dans la définition de l'axe : on peut considérer que c'est le cas de X_3 resp. X_4. On constate que sur le cercle des corrélations que l'axe principal oppose X_1 à X_2 et X_4 c'est à dire la latitude aux températures moyennes de Janvier et de Juillet.

Le 2ème axe oppose X_2 à X_5 donc la longitude et la hauteur annuelle des précipitations. C'est donc un axe qui va permettre de distinguer les villes en altitude au climat rude entière des villes plurieuses.

Il semblerait qu'il n'y ait pas de corrélation entre X_3 et X_5, et X_2 et X_4, car les variables sont quasiment orthogonales les unes par rapport aux autres.

B- Classification.

1. Soient A et B, $A \cap B = \emptyset$

$$D(A,B) = \max_{M \in A,\, M' \in B} d(M,M')$$

Montrons que : Si A, B, C trois parties disjointes de C, $D(A \cup B, C) = \max(D(A,C), D(B,C))$

$$D(A \cup B, C) = \max_{M \in A \cup B,\, M' \in C} d(M,M') \quad \text{or } A \cap B = \emptyset$$

$$= \max\left(\max_{M \in A,\, M' \in C} d(M,M'), \max_{M \in B, M' \in C} d(M,M')\right) \qquad M \in A \text{ ou } M \in B \text{ réel}$$

$$= \max(D(A,C), D(B,C))$$

2- critère d'aggrégation du lien maximum
algorithme de construction ascendante hiérarchique.

$P_{16} = (A_1, A_2, A_3, A_4, A_5)$

a) Continuons l'algorithme :
à l'étape 16 on a P_{16}, on construit la partition P_{17} en agrégeant A_ℓ^{16} et $A_{\ell'}^{16}$ vérifiant : $D(A_\ell^{16}, A_{\ell'}^{16}) = \min_{\ell \neq \ell'} D(A_\ell^{16}, A_{\ell'}^{16})$
avec D : distance du lien maximum,
d'après le tableau des distances, A_4 et A_3 vérifient cette propriété, c'est donc eux que l'on aggrège.

$P_{17} = (A_1, A_2, \{A_3, A_4\}, A_5)$

$\nu = 2,90$ tableau des distances.

A_1	0			
A_2	3,90	0		
A_{34}	4,06	6,00	0	
A_5	6,03	4,24	6,05	0
	A_1	A_2	A_{34}	A_5

$D(A_3, A_4) = \max_{M \in A_3, M' \in A_4} d(M, M')$

$D(A_3 \cup A_4, A_5) =$
$\max(D(A_3, A_5), D(A_4, A_5))$
$= 6,05$
On procède de même pour le reste.

On aggrège ensuite A_2 et A_1

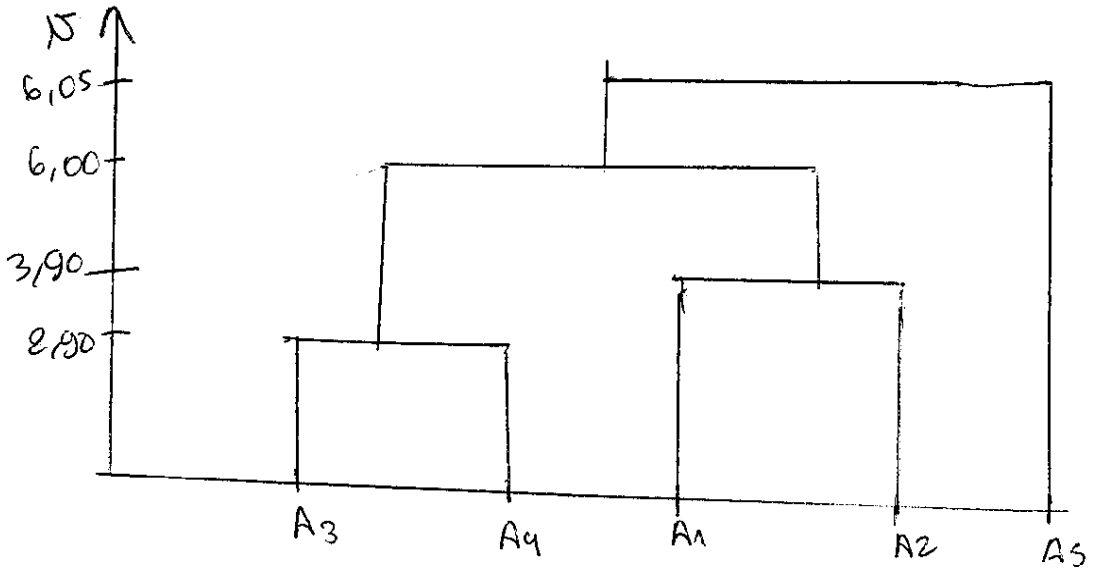
$P_{18} = (\{A_1, A_2\}, A_{34}, A_5)$ $\nu = 3,90$

A_{12}	0		
A_{34}	6,00	0	
A_5	6,03	6,05	0
	A_{12}	A_{34}	A_5

On aggrège ensuite A_{34} et A_{12} : $P_{19} = (\{A_{12}, A_{34}\}, A_5)$ $\nu = 6,00$

$D(A_{12} \cup A_{34}, A_5) = 6,05 \rightarrow P_{20} = (A_{12345})$ $\nu = 6,05$

partie du Dendrogramme :



b). On regarde le dendrogramme par le haut de manière à ce qu'on trouve 2 branches : on "coupe".

la classification en 2 classes obtenue est

A_5 et A_{1234} autrement dit :

$A_{F1} = \{$Athènes$\}$

$A_{F2} = \{$Lyon, Milan, Londres, Paris, Amsterdam, Dublin, Barcelone, Rome, Lisbonne, Madrid, Berlin, Vienne, Prague, Copenhague, Sofia, Varsovie, Helsinki, Stockholm, Oslo$\}$.

3 - partition en 2 classes :

$C_1 = \{Mo\}$ $C_2 = \{C\setminus\{Mo\}\}$

Ic l'inertie du nuage, W(P) critère de la somme des inertie pour $P = \{\{Mo\}, \{C\setminus\{Mo\}\}\}$.

a - Montrons que $W(P) = n\,Ic - \frac{M-2}{n-1}\|GMo\|^2$

b- $P_1 = \{ \{Athènes\}, \{C / \{Athènes\}\}$

$$W_1 = 20 \times Ic - \frac{20-2}{20-1} \| G M_0 \|^2$$

$$X_G = \frac{15 \times 0{,}80 + 5 \times -2{,}4}{20} = 0 \qquad Y_G = \frac{15 \times 0{,}13 + 5 \times -0{,}39}{20} = 0$$

Pour les données normalisées le centre de gravité est à l'origine et $tr(\Gamma) = tr(R) = p = Ic$

$$W_1 = 20 \times 5 - \frac{18}{19}\left[(2{,}75)^2 + (2{,}81)^2\right]$$
$$= 91{,}986$$

Athènes $(2{,}75 ; 2{,}81)$

— C — Discrimination.

$$I(B_2) = \frac{1}{n_{C_1}} \sum_{i, M_i \in C_1} d^2(C, M_i) = \frac{1}{n_{C_1}} \sum_{i, M_i \in C_1} \sum_{j=1}^{2} (x_{ij} - \bar{x}_j)^2$$

$$= \sum_{j=1}^{2} \left[\frac{1}{n_{C_1}} \sum_{i, M_i \in C_1} (x_{ij} - \bar{x}_j)^2\right] = \sum_{j=1}^{2} \sigma_j^2$$

$$= 0{,}99^2 + 1{,}04^2 = 2{,}06$$

$I(C_2) = 0,3^2 + 1,37^2 = 1,97$

$$W_2 = \sum_{l=1}^{k} n_2 I_l = m_1 I_1 + m_2 I_2 = 15 \times 2,06 + 5 \times 1,97$$
$$= 40,75$$

$W_1 > W_2$. La répartition obtenue en 1 est meilleure au sens du critère que celle ci (2).

- C - Discrimination (suite)

2- $\alpha = \frac{m_2 I_2}{m_1 I_1}$ $\qquad \overrightarrow{\Omega G_2} = \alpha \overrightarrow{\Omega G_1}$

$$r = \frac{\sqrt{\alpha}}{|1-\alpha|} d(G_1, G_2) = \frac{\sqrt{2}}{|1-\alpha|} \sqrt{(-2,4-0,80)^2 + (-0,39-0,13)^2}$$

$G_1 (0,80 ; 0,13)$ $\qquad G_2 (-2,4 ; -0,39)$

$$\overrightarrow{G_1 G_2} = G_1\vec{\Omega} + \vec{\Omega G_2} = -\vec{\Omega G_1} + \alpha \vec{\Omega G_1} = (\alpha - 1) \vec{\Omega G_1}$$

$$\overrightarrow{G_1 G_2} = \begin{pmatrix} -2,4 - 0,80 \\ -0,39 - 0,13 \end{pmatrix} = \begin{pmatrix} -3,2 \\ -0,52 \end{pmatrix}$$

$$(\alpha - 1) \vec{\Omega G_1} = (\alpha - 1) \begin{pmatrix} 0,81 - x_\Omega \\ 0,13 - y_\Omega \end{pmatrix}$$

$$\Rightarrow \begin{cases} -3,2 = (\alpha - 1)(0,81 - x_\Omega) \\ -0,52 = (\alpha - 1)(0,13 - y_\Omega) \end{cases}$$

$$\Rightarrow \begin{cases} x_\Omega = 0,81 + \frac{3,2}{\alpha - 1} \\ y_\Omega = 0,13 + \frac{0,52}{\alpha - 1} \end{cases}$$

A.N : $\alpha = \frac{m_2 I_2}{m_1 I_1} = \frac{5 \times 1,97}{15 \times 2,06} = 0,32$

$r = 2,7$

$\Omega (x_\Omega, y_\Omega) = \Omega (-3,89 ; -0,63)$

b - La ville d'Athènes :

$d^2(\Omega, \text{Athènes}) = (-3,89 + 3,75)^2 + (-0,63 + 2,81)^2 = 6,052$

comparons le résultat avec $r^2 = 7,29 \Rightarrow d^2(\Omega, \text{Athènes}) < r^2$

Athènes est situé à l'intérieur du cercle

donc classe : C2 . Athènes est bien classée.

c - $M_{\text{Toulouse}} [43,6 ; 1,43 \quad 4,7 \quad 20,9 \quad 656]$

projection dans le plan principal :

on normalise M Toulouse :

$$M_T' \left[\frac{43,6 - 48,93}{7,01} \quad \frac{1,43 - 9,77}{10,12} \quad \frac{4,7 - 1,99}{4,94} \right.$$

$$\left. \frac{20,9 - 20,4}{3,32} \quad \frac{656 - 639}{138} \right]$$

$$M_T' [-0,76 \ ; -0,82 \ ; 0,55 \ ; 0,15 \ ; 0,12]$$

$\Pi_1(M_T) = -0,76 \times 0,5780 \quad -0,82 \times 0,2676 + 0,55 \times -0,5691$

$+ 0,15 \times -0,5177 + 0,12 \times 0,0489$

$= -1,043$

$\Pi_2(M_T) = -0,76 \times 0,1116 - 0,82 \times -0,6700 + 0,55 \times 0,1509 + 0,15 \times -0,3273$

$+ 0,12 \times 0,6393 = 0,57 \rightarrow$ calcul de $d^2(\Omega, M_T)$

$$d^2(\Omega, M_7) = (-3,89 + 1,043)^2 + (-0,63 - 0,57)^2$$

$$= 9,54 > 7,29 = r^2$$

en dehors du cercle, donc doit être affecté à la classe C_1.

Groupe ESIEE **06/11/08**

O.de Cambry **MA-412**

Traitement statistique des données

Examen

Durée :**2 heures** Sujet à traiter **avec** documents

On considère le tableau de données noté $\mathcal{C}$ suivant concernant 12 aliments.

Numéro	Aliment	X_1	X_2	X_3	X_4
1	céréales	13.0	71.0	12.0	2.5
2	pommes de terre	80.0	17.0	2.0	0
3	haricots secs	11.0	61.0	22.0	1.5
4	epinards	90.0	5.5	3.0	0.3
5	chou	92.0	5.5	1.5	0.2
6	orange	86.0	12.0	1.0	0.2
7	poulet	73.5	0	22.0	4.0
8	porc	56.0	0	16.0	27.0
9	cabillaud	81.0	0	17.5	0.5
10	oeuf	74.0	1.0	13.0	11.5
11	yoghourt	88.0	8.0	3.5	2.5
12	camembert	52.0	4.0	17.5	24.5
	Moyenne	66.38	15.42	10.92	6.23
	Ecart-type	27.09	23.25	7.89	9.25

Les variables étudiées sont des quantités de composants fondamentaux dans chacun des aliments

- X_1 :eau
- X_2 :glucides
- X_3 :protides
- X_4 :lipides

- A - Analyse en composantes prindipales

On a effectué une Analyse en Composantes Principales (ACP) sur les données normalisées, dont les résultats sont rassemblés à la suite des questions.

1. *Représenter et commenter la boite à moustaches pour* X_1.
2. *Compléter le cercle des corrélations en ajoutant* X_1 *et* X_2 *(Figure 1).*
3. *Commenter les résultats de L'ACP.*

- B - Classification

Dans la suite on travaille sur les données normalisées. Les distances entre les points se trouvent en annexe.

1. Soient A et B, 2 groupements de points disjoints, on définit le critère du lien minimum entre ces 2 groupements par : $D(A,B) = \min_{M\in A, M'\in B} d(M,M')$.

   La hiérarchie $\mathcal{H}(\mathcal{C})$ du saut minimum définie à partir du critère d'aggrégation du lien minimum a donné le dendrogramme représenté sur la Figure 2.

   (a) *Déterminer les indices d'aggrégation et compléter le dendrogramme.*

   (b) *Représenter sur la Figure 3 l'arbre couvrant minimal associé à cette hiérarchie en notant sur les arètes de l'arbre les indices d'aggrégation correspondants.*

2. Dans la suite on notera pour $A \in \mathcal{H}(\mathcal{C})$, $v(A)$ l'indice d'aggrégation associé à la constitution de l'ensemble A dans la hiérarchie. On considère à présent l'application Δ définie pour 2 points M et M' de $\mathcal{C}$ par $\Delta(M,M') = \min_{A\in\mathcal{H}(\mathcal{C})} \{v(A); M \in A, M' \in A\}$

   (a) *Montrer que Δ est une distance ultramétrique.*

   (b) *Déterminer Δ(orange,céréales) et Δ(camembert,yoghourt).*

- C - Discrimination

On considère maintenant la partition en 2 classes des données dans le plan principal définie ci-dessous :

- C_1={cereales, pommes de terre, haricots secs, epinards, chou, orange, yoghourt}
- C_2={poulet, porc, cabillaud, oeuf, yoghourt, camembert}

On notera n_l le nombre d'élément de la classe C_l. Dans la suite on souhaite classer dans C_1 ou C_2 le fromage frais dont les caractéristiques sont :
eau=60, glucides=4, protides=13, lipides=4.

1. *Déterminer les coordonnées dans le plan principal du* fromage frais *et placer le point correspondant sur la figure 4.*

2. On souhaite discriminer les 2 classes C_1 et C_2 en utilisant la méthode des 2 plus proches voisins.
   Déterminer la valeur des fonctions discriminantes des 2-PPV pour le fromage frais *et en déduire la classe dans laquelle il est affecté.*

3. On utilise maintenant la méthode de Bayes dite naive dans laquelle la matrice de covariance est diagonale et identique pour les 2 classes. On dispose donc des 2 probabilité a priori α_1 et α_2 et on suppose que sachant qu'un individu est dans la classe C_i sa position suit une loi $\mathcal{N}(m_i, V)$ où $V = \begin{bmatrix} s_1^2 & 0 \\ 0 & s_2^2 \end{bmatrix}$.

   On rappelle qu'une fonction discriminante est linéaire si pour le point $M(x_1,x_2)$ la décision est de la forme $\delta(x) = \left| \begin{array}{lll} 1 & \text{si} & L(M) \leq 0 \\ 2 & \text{si} & L(M) \geq 0 \end{array} \right.$ avec $L(M) = a_1x_1 + a_2x_2 + b$.

   (a) *Vérifier que la fonction discriminante est linéaire.*

   (b) *Déterminer les estimations de α_1, α_2, m_1 et m_2.*

(c) Si on note $\sigma_i^2(l)$ la variance sur l'axe u_i pour les éléments de la classe C_l, on estime les coefficient de la matrice V par $s_i^2 = \dfrac{(n_1-1)\sigma_i^2(1)+(n_2-1)\sigma_i^2(2)}{n_1+n_2-2}$.
En déduire V.

(d) On a obtenu $a_1 = 0.55$, $a_2 = 5.5$, $b = -1.1$.
Tracer la droite séparatrice correspondante sur la Figure 4 et déterminer l'affectation de fromage frais

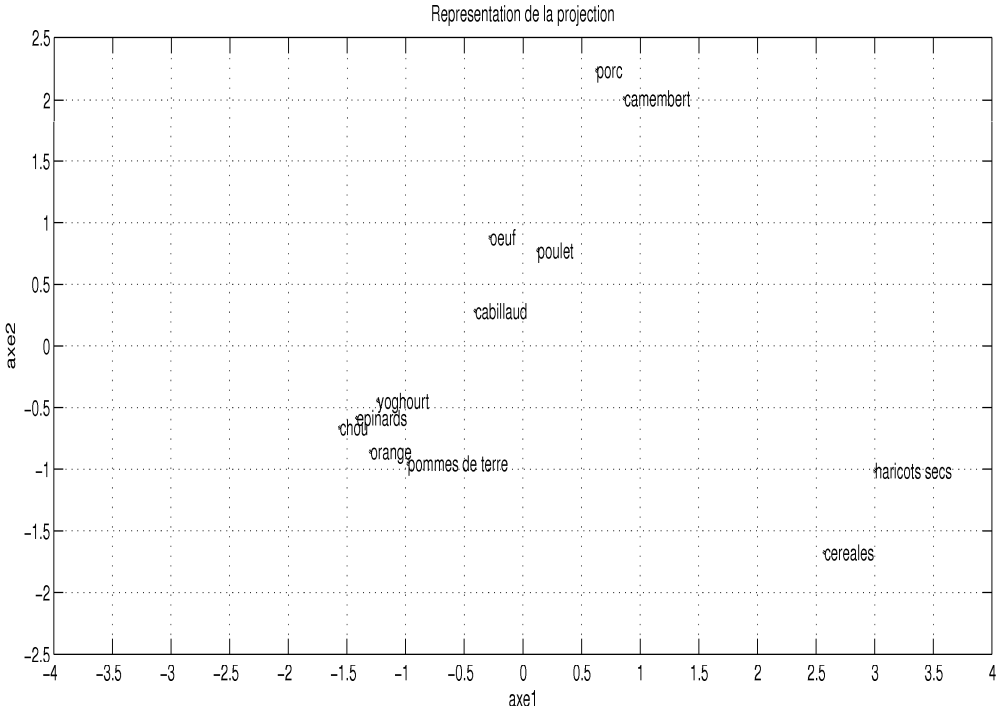
Barème indicatif :

- A - : 5pts

- B - : 7pts

- C - : 8pts

Annexe



- A -

Matrice de corrélation : $R = \begin{pmatrix} 1.0000 & -0.8346 & -0.5972 & -0.2178 \\ -0.8346 & 1.0000 & 0.1617 & -0.2898 \\ -0.5972 & 0.1617 & 1.0000 & 0.4160 \\ -0.2178 & -0.2898 & 0.4160 & 1.0000 \end{pmatrix}$

Valeurs propres :

$$\begin{aligned}\lambda_1 &= 2.137\\ \lambda_2 &= 1.414\\ \lambda_3 &= 0.448\\ \lambda_4 &= 0.001\end{aligned}$$

Vecteurs propres :

$$U = \begin{pmatrix} -0.6746 & 0.0859 & -0.1931 & 0.7073 \\ 0.5201 & -0.5290 & 0.2406 & 0.6259 \\ 0.4930 & 0.4107 & -0.7348 & 0.2198 \\ 0.1770 & 0.7377 & 0.6041 & 0.2442 \end{pmatrix}$$

- B -

Distance entre les points

céréales	1	0											
pommes de terre	2	3.63	0										
haricots secs	3	1.34	4.06	0									
epinards	4	4.17	0.63	4.47	0								
chou	5	4.28	0.67	4.63	0.20	0							
orange	6	3.96	0.33	4.38	0.41	0.36	0						
poulet	7	3.99	2.68	3.50	2.53	2.73	2.78	0					
porc	8	4.37	3.60	4.22	3.56	3.69	3.67	2.68	0				
cabillaud	9	4.02	2.10	3.73	1.88	2.08	2.16	0.74	3.01	0			
oeuf	10	3.88	2.00	3.81	1.86	2.02	2.05	1.40	1.84	1.34	0		
yoghourt	11	4.02	0.59	4.33	0.28	0.40	0.44	2.43	3.32	1.84	1.66	0	
camembert	12	4.06	3.50	3.85	3.49	3.63	3.60	2.43	0.40	2.81	1.72	3.25	0
		1	2	3	4	5	6	7	8	9	10	11	12

- C -

Coordonnées dans le plan principal et caractéristiques des classes

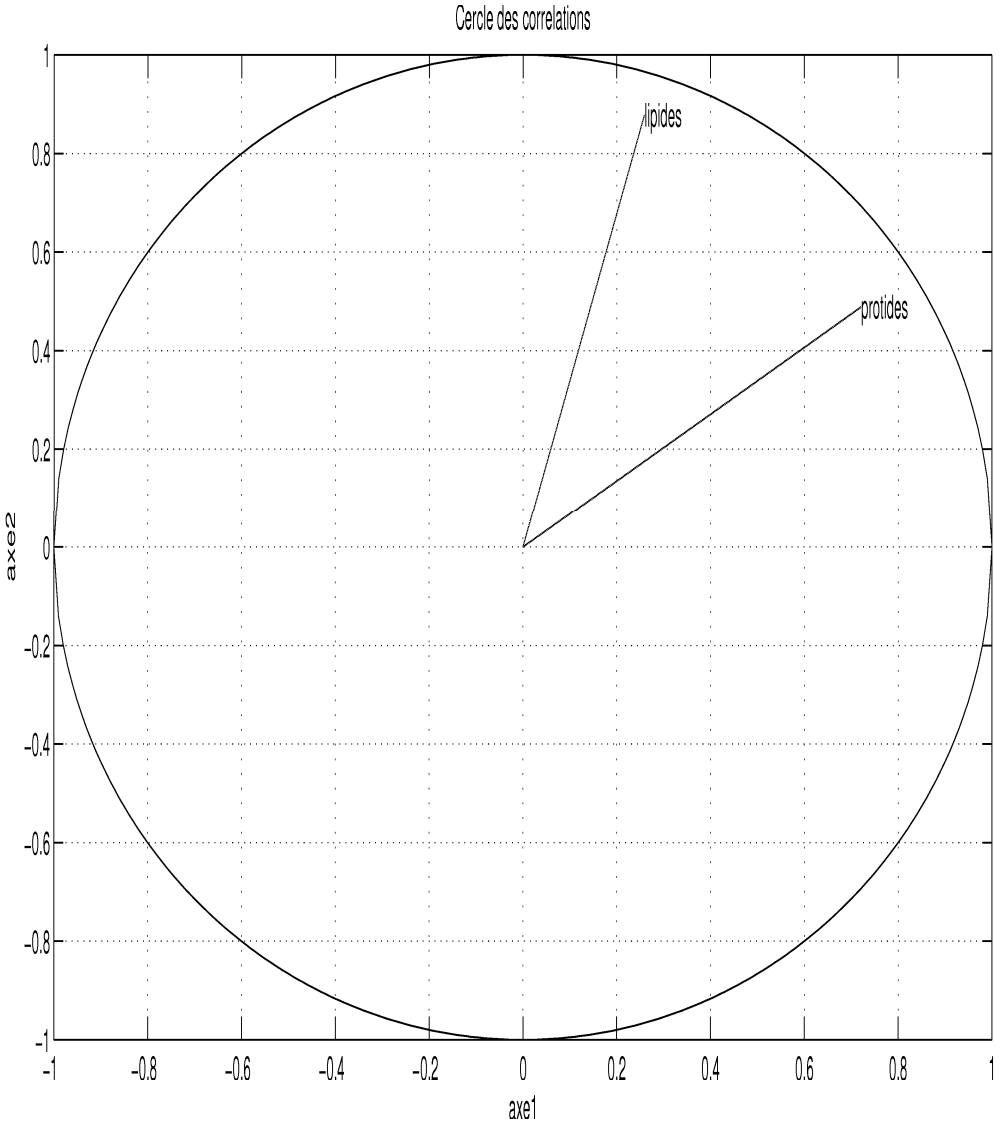
d : distance entre l'aliment et fromage frais dans le plan principal.
Xp_1, Xp_2 : coordonnées du point dans le plan principal

C_1	Nom	Xp_1	Xp_2	d	C_2	Nom	Xp_1	Xp_2	d
1	céréale	2.57	-1.67	3.17	7	poulet	0.13	0.77	0.62
2	pommes de terre	-0.98	-0.96	1.48	8	porc	0.63	2.24	2.16
3	haricots secs	3.00	-1.01	3.23	9	cabillaud	0.41	0.28	0.41
4	epinards	-1.42	-0.58	1.60	10	oeuf	0.28	0.88	0.76
5	chou	-1.56	-0.66	1.76	12	camembert	0.86	2.01	2.04
6	orange	-1.30	-0.86	1.65					
11	yoghourt	-1.24	-0.45	1.38					
	Moyenne	-0.13	0.88			Moyenne	0.19	1.24	
	Ecart-type	1.86	0.37			Ecart-type	0.50	0.76	

Ne pas oublier de rendre cette page et les suivantes avec la copie

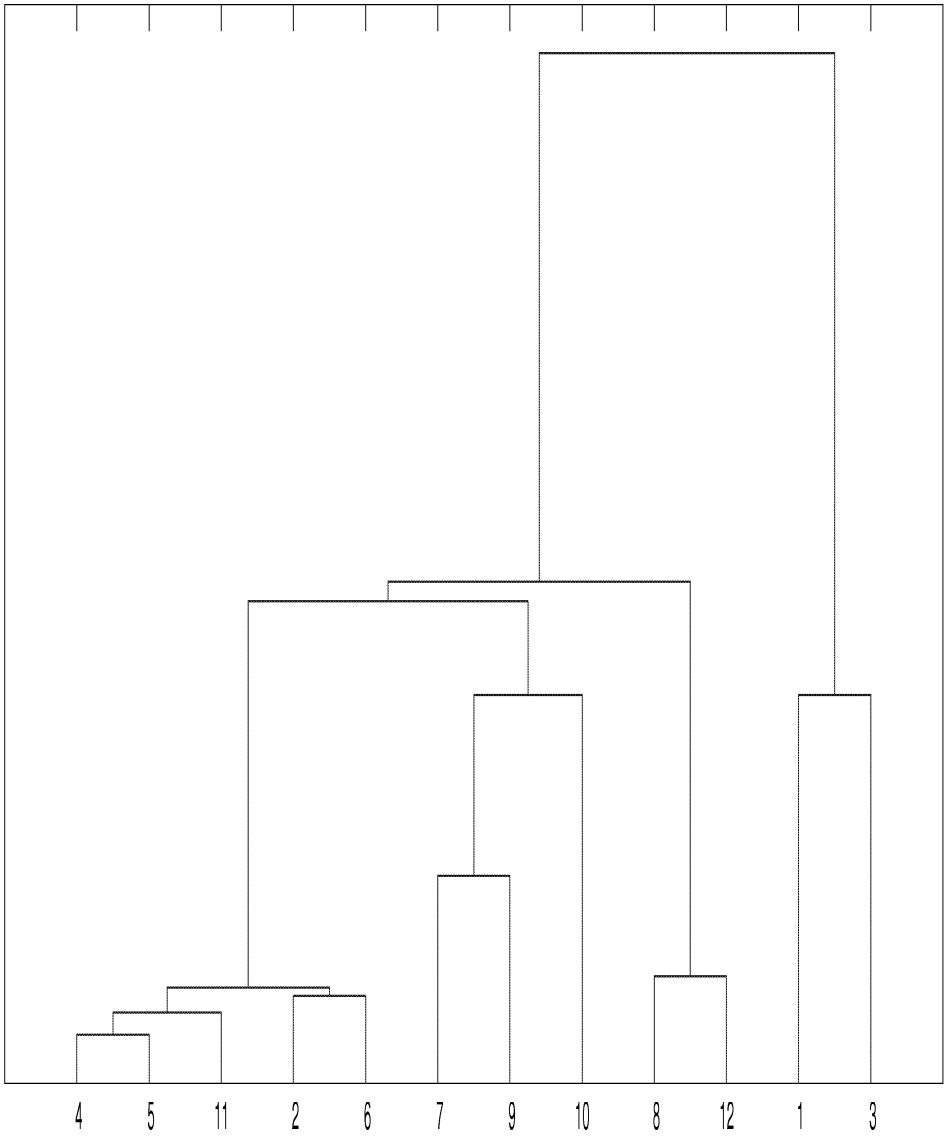
NOM :

Question - A - 2 : Figure 1



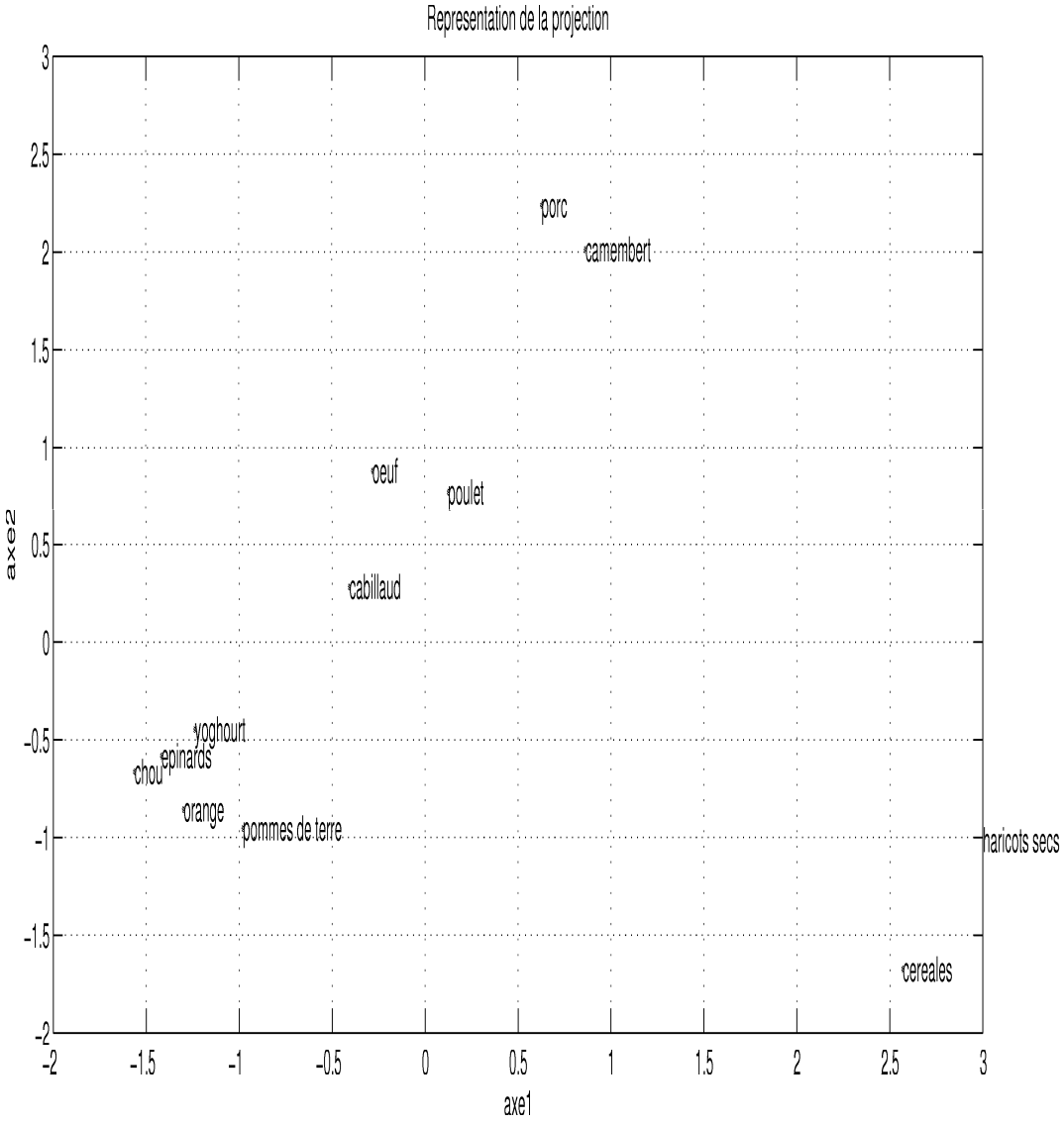
NOM :

Question - B - 1 - a : Figure 2



NOM :

Question - B - 1 - b : Figure 3



NOM :

Question - C - : Figure 4