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

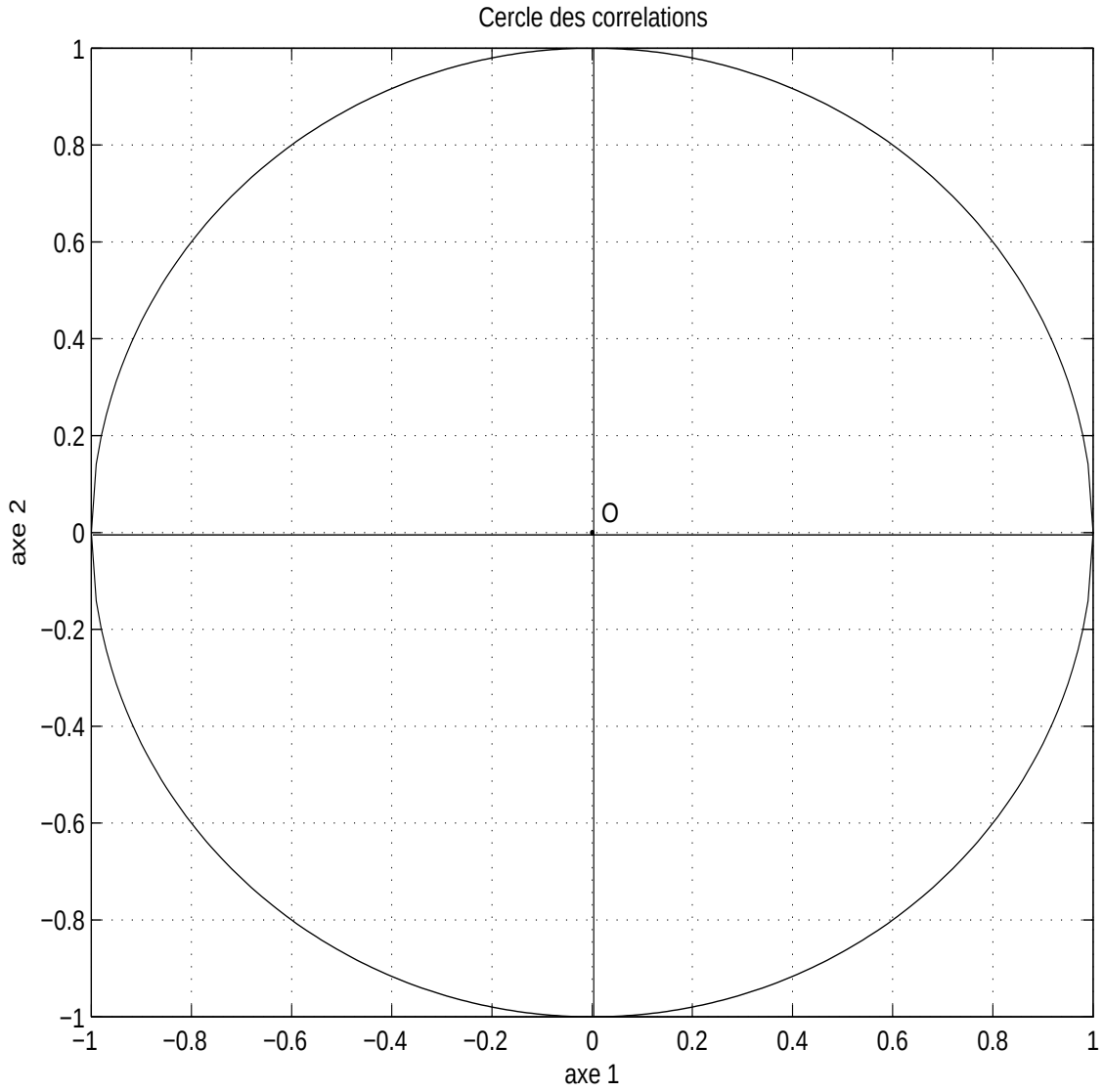

FIG. 1 – Cercle de corrélation

Question - B - 1

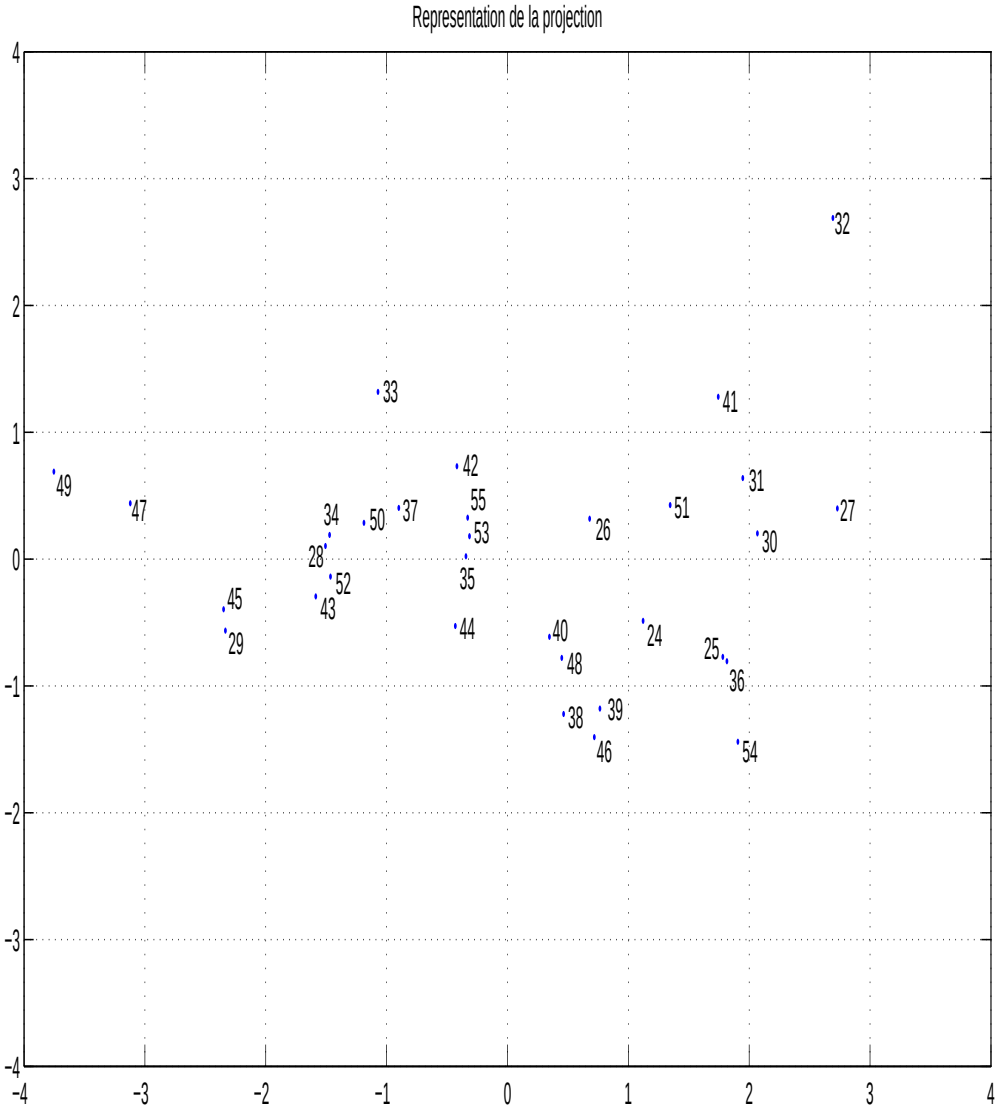

FIG. 2 – Projection sur le plan principal