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

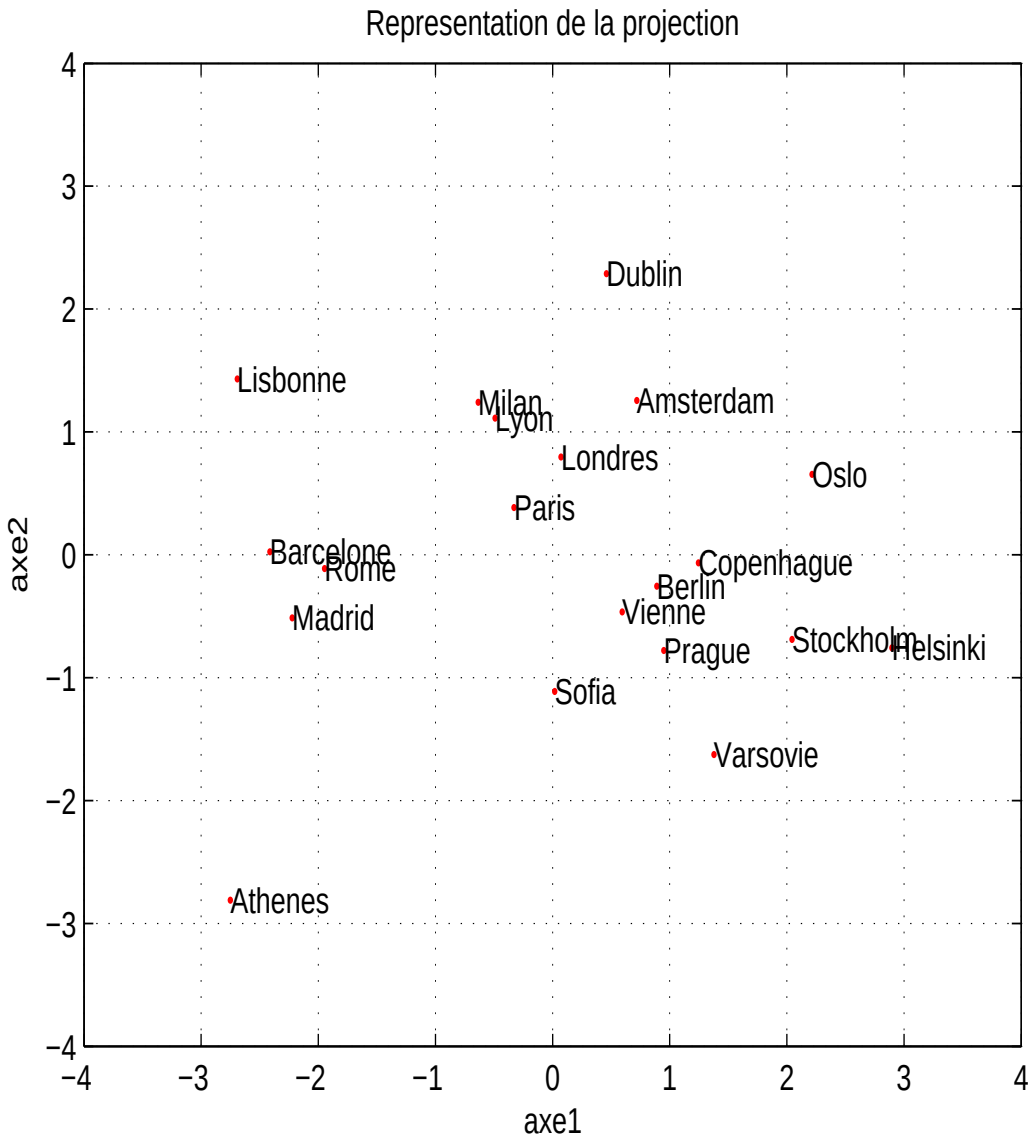
$$R = \begin{pmatrix} 1.0000 & 0.2605 & -0.8004 & -0.8653 & 0.0787 \\ 0.2605 & 1.0000 & -0.5796 & 0.0302 & -0.2521 \\ -0.8004 & -0.5796 & 1.0000 & 0.6605 & -0.0487 \\ -0.8653 & 0.0302 & 0.6605 & 1.0000 & -0.1486 \\ 0.0787 & -0.2521 & -0.0487 & -0.1486 & 1.0000 \end{pmatrix}$$

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

- 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

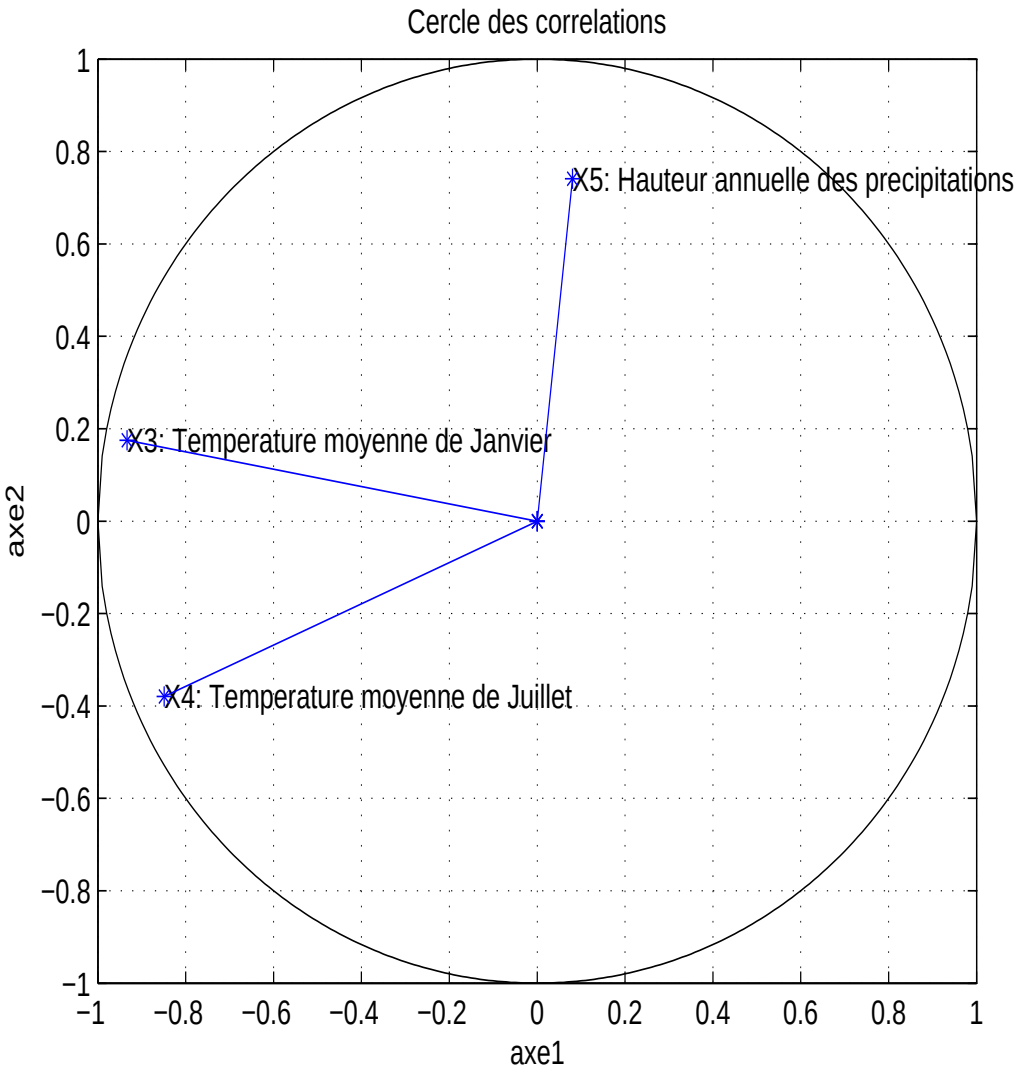


FIG. 1 – Cercle de corrélation

Ne pas oublier de rendre cette page avec la copie

Figure 1

NOM :

PRENOM :

Question - B - 2

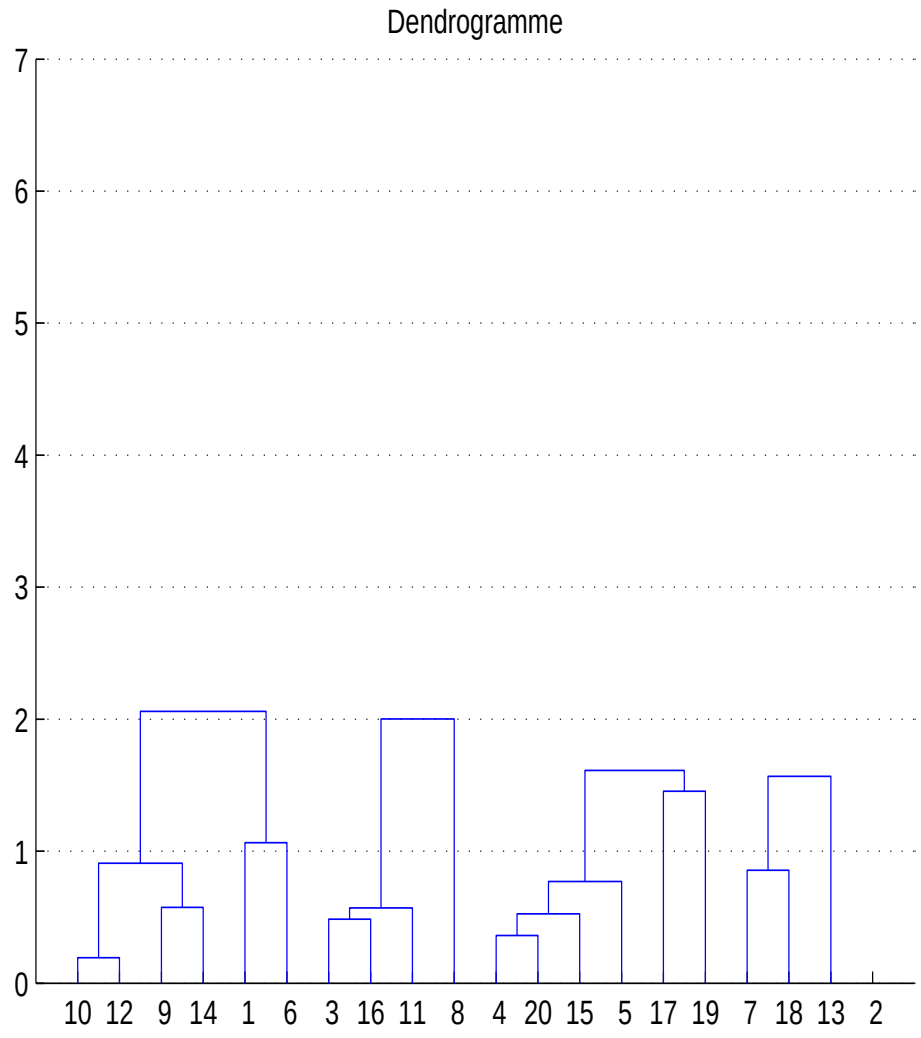


FIG. 2 – Projection sur l'axe principal