Groupe ESIEE **09/11/10**

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 $\mathcal{C}$, se trouvant en annexe, concernant 12 modèles d'automobile.
Les variables étudiées sont des caractéristiques techniques de ces voitures.

- A - Analyse des données

1. *Représenter la boite à moustache et un histogramme en 4 classes pour la hauteur (X_3).*
2. On a effectué une Analyse en Composantes Principales (ACP) sur les données normalisées, dont les résultats sont rassemblés à la suite des questions.

   (a) *Compléter le cercle des corrélations en Figure 1.*
   (b) *Commenter les résultats de L'ACP.*

- B - Classification

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

1. On considère une classification en 3 classes par la méthode des kmeans en 2 phases :
   phase1 : algorithme des centres mobiles
   phase2 : algorithme de transfert
   A l'issue de l'étape m l'algorithme est toujours dans la phase 1 et la classification suivante est obtenue :
   $L_m = [2 \quad 1 \quad 1 \quad 2 \quad 2 \quad 1 \quad 1 \quad 2 \quad 3 \quad 1 \quad 2 \quad 2]$
   Les éléments concernant cette partition sont donnés en annexe.
   A l'issue de l'algorithme on obtient la classification suivante :
   $L_f = [2 \quad 3 \quad 1 \quad 2 \quad 2 \quad 1 \quad 1 \quad 2 \quad 3 \quad 1 \quad 2 \quad 2]$
   (a) *Décrire la fin de l'algorithme.*
   (b) *Calculer la valeur finale du critère de la somme des inerties.*
2. On souhaite réaliser la classification hiérarchique pour la distance du lien minimum de ces données.
   (a) *Déterminer et représenter sur la Figure 2 l'arbre couvrant minimal en utilisant l'algorithme de Prim .*
   (b) *En déduire le dendrogramme et la classification en 3 classes obtenue.*

- C - Discrimination

On considère maintenant la partition en 3 classes L_f obtenue à la question B - 1.

1. Soit $\mathcal{C}$ un nuage de point de $\mathbb{R}^p$ et V, la matrice de covariance de ce nuage.
   On définit la mesure de voisinage de la distance au centre pondérée par la covariance entre un point M et une classe A d'éléments de $\mathcal{C}$ par :

   $S_A(M) = (M - G_A)^t V^{-1}(M - G_A)$ où G_A représente le centre de gravité de A.

   On souhaite utiliser cette mesure de voisinage pour discriminer les classes.

   (a) *Déterminer l'équation de la surface séparatrice entre 2 classes A_1 et A_2 en fonction de leurs centres de gravités G_1 et G_2 et de V.*

   (b) *Quelle est le type de la courbe séparatrice lorsque $p = 2$.*

2. On suppose que pour les données sur les automobiles, projetées dans le plan principal, la matrice de covariance V est diagonale .

   (a) *Déterminer l'expression de $S_A(M)$ en fonction de $M(x_1; x_2)$, $G_A(\bar{x}_1; \bar{x}_2)$ et* $V = \begin{bmatrix} \gamma_1 & 0 \\ 0 & \gamma_2 \end{bmatrix}$.

   (b) *Déterminer sans calcul γ_1 et γ_2 à partir des données en annexe.*

3. On souhaite déterminer la classe du modèle *Fiat Croma* dont les caractéristiques sont :

   - X_1 :Cylindrée : 2200 cm^3
   - X_2 :Longueur : 475 cm
   - X_3 :Hauteur : 159 cm
   - X_4 :Poids : 1430 kg

   (a) *Calculer les coordonnées du point représentatif de ce modèle dans le plan principal et ajouter le dans la Figure 2.*

   (b) *Calculer $S_{A_i}(M)$ pour i=1,2,3.*

   (c) *En déduire la classe à laquelle la Fiat Croma doit être affectée en utilisant la méthode de la question - 1 -.*

Barème indicatif :

- A - : 6pts

- B - : 7pts

- C - : 7pts

Annexes

Données

	Marque	Modèle	X_1	X_2	X_3	X_4
1	Citroen	C3	1587	393	156	1090
2	Citroen	C4 Picasso	1997	447	166	1511
3	Citroen	C5	2946	462	148	1567
4	Peugeot	207	1360	403	147	1640
5	Peugeot	307	1997	420	151	1282
6	Peugeot	407	2230	468	145	1398
7	Peugeot	607	2946	487	146	1625
8	Renault	Clio	1390	381	142	1019
9	Renault	Espace	3498	466	173	1936
10	Renault	Laguna	1998	458	143	1490
11	Renault	Megane	1998	421	146	1165
12	Renault	Twingo	1149	343	142	872
	Moyenne		2091	429	150	1383
	Ecart-type		717	43.0	9.85	306

- X_1 :Cylindrée(cm^3)
- X_2 :Longueur(cm)
- X_3 :Hauteur(cm)
- X_4 :Poids(kg)

- A - Analyse des données

Matrice de corrélation :

$$R = \begin{pmatrix} 1.0000 & 0.8380 & 0.4924 & 0.7655 \\ 0.8380 & 1.0000 & 0.3006 & 0.7915 \\ 0.4924 & 0.3006 & 1.0000 & 0.5386 \\ 0.7655 & 0.7915 & 0.5386 & 1.0000 \end{pmatrix}$$

Valeurs propres :

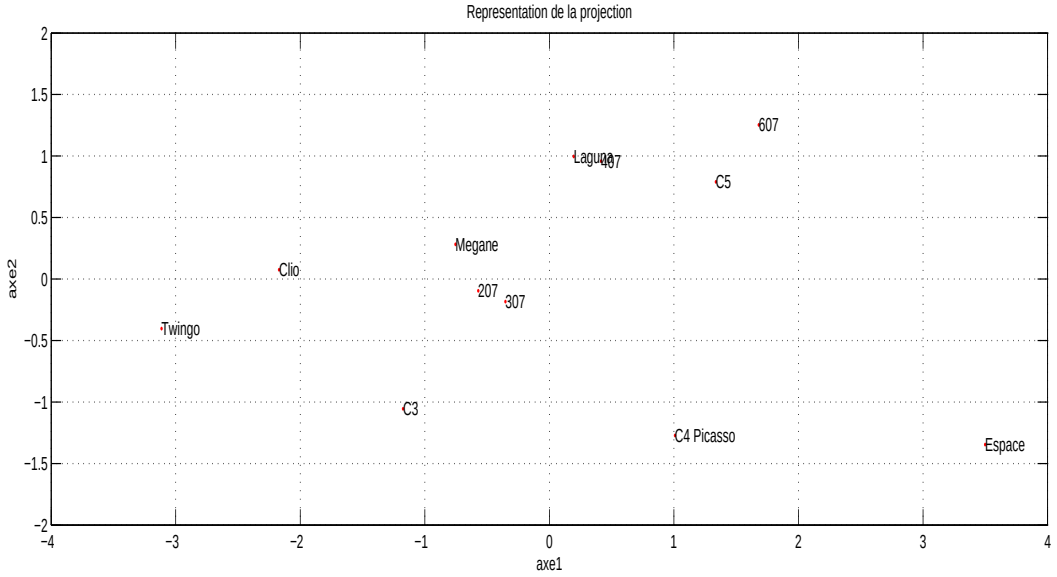
$$\begin{aligned} \lambda_1 &= 2.91 \\ \lambda_2 &= 0.75 \\ \lambda_3 &= 0.23 \\ \lambda_4 &= 0.11 \end{aligned}$$

Vecteurs propres :

$$U = \begin{pmatrix} 0.5430 & 0.1518 & 0.6327 & 0.5308 \\ 0.5219 & 0.4436 & 0.0551 & -0.7265 \\ 0.3750 & -0.8827 & 0.1092 & -0.2613 \\ 0.5404 & 0.0316 & -0.7647 & 0.3495 \end{pmatrix}$$

Coordonnées des points dans le plan principal :

	Modèle	C_1	C_2
1	C3	-1.17	-1.05
2	C4 Picasso	1.01	-1.27
3	C5	1.34	0.79
4	207	-0.57	-0.10
5	307	-0.35	-0.18
6	407	0.42	0.96
7	607	1.68	1.25
8	Clio	-2.17	0.08
9	Espace	3.50	-1.35
10	Laguna	0.20	0.10
11	Megane	-0.75	0.28
12	Twingo	-3.11	-0.40



- B - Classification

B - 1 : Kmeans

Partition à l'étape m $L_m = [2 \quad 1 \quad 1 \quad 2 \quad 2 \quad 1 \quad 1 \quad 2 \quad 3 \quad 1 \quad 2 \quad 2]$

Centres de gravités à l'étape m :

$$\begin{aligned} G_1^m &= (0.93;\ 0.55) \\ G_2^m &= (-1.36;\ -0.23) \\ G_3^m &= (3.50;\ -1.34) \end{aligned}$$

Distances aux centres des classes :

	Année	$d(M, G_1^m)$	$d(M, G_2^m)$	$d(M, G_3^m)$
1	C3	2.64	0.84	4.68
2	C4 Picasso	1.82	2.58	2.49
3	C5	0.48	2.88	3.04
4	207	1.63	0.80	4.26
5	307	1.47	1.00	4.02
6	407	0.66	2.13	3.85
7	607	1.03	3.38	3.17
8	Clio	3.13	0.87	5.84
9	Espace	3.19	4.98	0
10	Laguna	0.86	1.98	4.05
11	Megane	1.70	0.79	4.55
12	Twingo	4.15	1.77	6.68

Somme des inerties $W = 12.64$

Partition à l'étape finale $L_f = [2 \quad 3 \quad 1 \quad 2 \quad 2 \quad 1 \quad 1 \quad 2 \quad 3 \quad 1 \quad 2 \quad 2]$

Centres de gravités à l'étape finale :

$$\begin{aligned} G_1^f &= (0.91;\ 1) \\ G_2^f &= (-1.36;\ -0.23) \\ G_3^f &= (2.25;\ -1.31) \end{aligned}$$

B - 2 : Classification hiérarchique

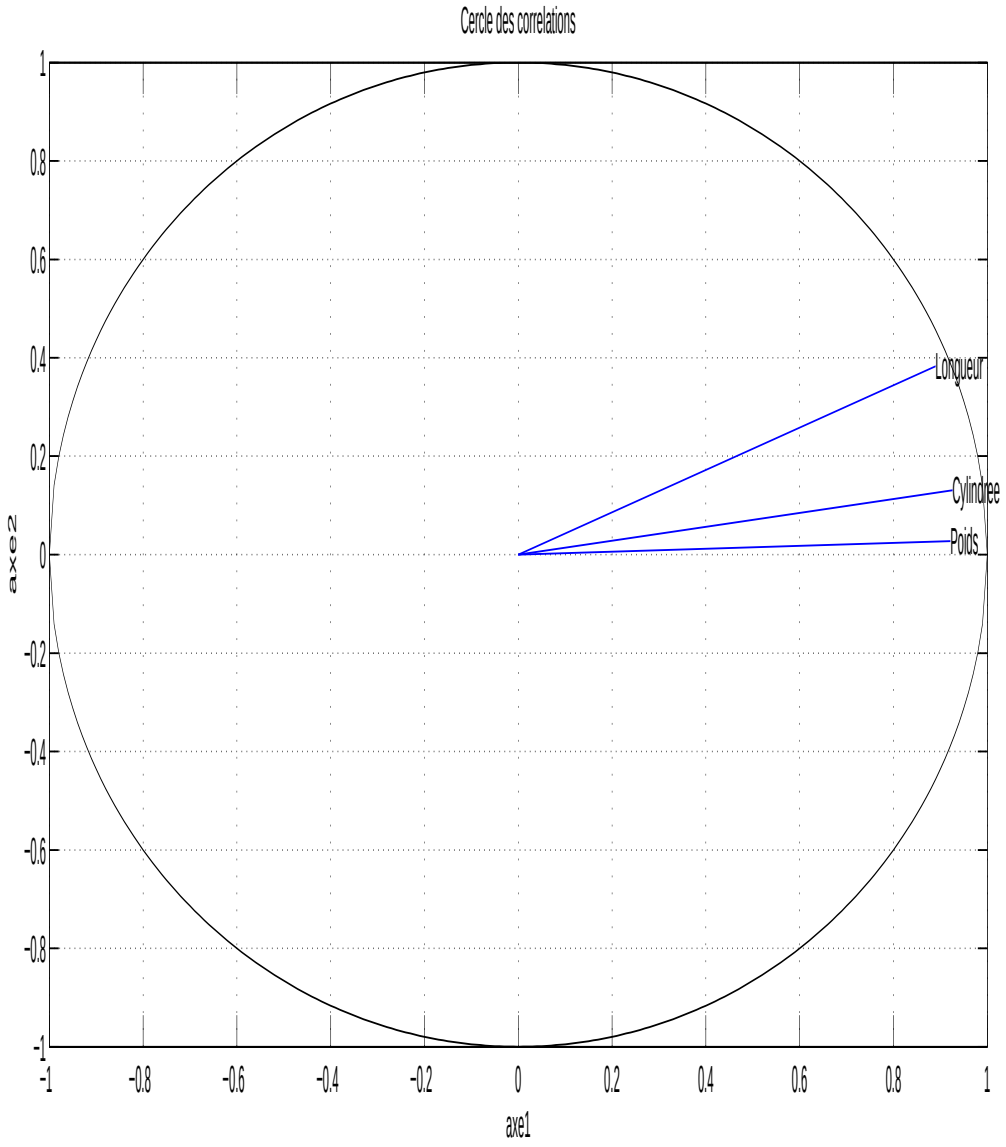
Tableau des distances entre les points

2	2.19										
3	3.11	2.08									
4	1.13	1.96	2.10								
5	1.19	1.74	1.95	0.23							
6	2.56	2.30	0.93	1.44	1.37						
7	3.67	2.61	0.57	2.62	2.49	1.29					
8	1.50	3.45	3.57	1.60	1.83	2.73	4.02				
9	4.68	2.49	3.03	4.25	4.02	3.84	3.17	5.84			
10	2.46	2.41	1.15	1.33	1.30	0.22	1.50	2.53	4.04		
11	1.40	2.35	2.15	0.41	0.61	1.35	2.62	1.43	4.55	1.18	
12	2.04	4.21	4.60	2.56	2.76	3.78	5.07	1.05	6.67	3.59	2.45
	1	2	3	4	5	6	7	8	9	10	11

Ne pas oublier de rendre cette page avec la copie

NOM :

Question - A - 2 : Figure 1



NOM :

Question - B - 2 et - C - 2 : Figure 2