INTRODUCTION NUMÉRIQUE À L'ACP

On a noté de 1 à 20 dix articles suivant quatre critères A, B, C, D :

n°	A	B	C	D
1	10	8	7	15
2	14	16	4	6
3	9	13	12	6
4	10	10	10	10
5	16	14	6	4
6	4	13	7	16
7	5	15	14	6
8	5	7	16	12
9	14	8	8	10
10	11	9	11	9

Tableau de données initiales

Les lignes du tableau sont les *unités statistiques* appelées aussi *individus statistiques*. On les note i = 1, …10. Les colonnes sont les *variables*, notées j = 1, … p. On a donc, dans le cas général, n lignes et p colonnes. Dans la pratique, le tableau peut avoir 1000 lignes et 50 colonnes.

1) Calculer les moyennes des notes pour chaque article. Quelle conclusion sur les articles peut-on tirer de ces moyennes ? Commenter le tableau des données à l'aide des résultats statistiques données ci-dessous.

critères	moyenne	écart-type	variance
A	9.8	3.94462	15.56
B	11.3	3.1	9.61
C	9.5	3.58469	12.85
D	9.4	3.82623	14.64

Moyennes et variances des notes données aux articles pour chaque critère

	A	B	C	D
A	1.000	0.087	-0.636	-0.505
B	0.087	1.000	-0.337	-0.583
C	0.636	-0.337	1.000	-0.007
D	0.505	-0.583	-0.007	1.000

Corrélations entre les notes données aux articles pour chaque critère

2) On considère les critères A et B. Représenter graphiquement les dix articles en fonction de leurs notes suivant ces deux critères. Quels sont les deux articles les plus différents ? Les plus proches ? Calculer les distances correspondantes. On considère maintenant les quatre critères A, B, C et D. Calculer les distances entre les articles 5 et 8, et entre 4 et 10 en tenant compte des quatre critères. Quelles sont les variables dont l'importance est la plus grande en moyenne dans le calcul de ces distances ? Quelle est finalement la transformation des données qu'il faut effectuer pour obtenir des distances indépendantes des coefficients et des variances ?

3) On note a, b, c, d les variables centrées réduites déduites de A, B, C, D. Compléter en utilisant les propriétés des variables centrées réduites le tableau des valeurs centrées réduites ci-dessous :

	a	b	c	d

1	0.0507	-1.0645	-0.6974	1.4636
2	1.0647	1.5161	-1.5343	-0.8886
3	-0.2028	0.5484	0.6974	-0.8886
4	0.0507	-0.4194	0.1395	0.1568
5	1.5718	0.8710	-0.9764	-1.4113
6	-1.4704	0.5484	-0.6974	1.7249
7	-1.2168	1.1935	1.2553	-0.8886
8	-1.2168	-1.3871	1.8133	0.6795
9	1.0647	-1.0645	-0.4184	0.1568
10	0.3042	-0.7419	0.4184	-0.1045

Compléter le tableau des distances entre les articles calculées à l'aide des valeurs centrées réduites précédentes.

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
1	0.000	13.921	10.144	2.824	14.403	4.983	16.052	8.629	2.814	3.872
2	13.921	0.000	7.524	8.669	1.258	14.894	13.092	27.300	8.998	10.105
3	10.144	7.524	0.000	2.405	6.328	10.383	1.756	8.479	6.546	2.615
4	2.824	8.669	2.405	0.000	7.683	6.410	6.546	5.618	1.756	0.314
5	14.403	1.258	6.328	7.683	0.000	19.272	13.134	25.029	6.773	7.861
6	4.983	14.894	10.383	6.410	19.272	0.000	11.124	11.207	11.565	9.406
7	16.052	13.092	1.756	6.546	13.134	11.124	0.000	9.430	14.199	7.375
8	8.629	27.300	8.479	5.618	25.029	11.207	9.430	0.000	10.563	5.290
9	2.814	8.998	6.546	1.756	6.773	11.565	14.199	10.563	0.000	1.451
10	3.872	10.105	2.615	0.314	7.861	9.406	7.375	5.290	1.451	0.000

Carrés des distances entre les dix articles

On vérifiera (ou on admettra) que la somme des carrés des distances est égale à 800 ($2 \times n^2 \times p$).

4) On donne maintenant les variables C_1 et C_2 définies par les coefficients appliqués aux variables centrées réduites a, b, c, d ci-dessous :

$$C_1 = -0.5523 \times a - 0.4662 \times b + 0.4687 \times c + 0.5079 \times d$$
$$C_2 = -0.3941 \times a + 0.4783 \times b + 0.5820 \times c - 0.5265 \times d$$
$$C_3 = -0.4942 \times a + 0.6094 \times b - 0.4445 \times c + 0.4322 \times d$$
$$C_4 = 0.5436 \times a + 0.4272 \times b + 0.4940 \times c + 0.5273 \times d$$

Les variables C_1, C_2, C_3, C_4 sont les *composantes principales*. Leurs variances sont appelées *valeurs propres*. Chaque composante principale est de variance maximale sous la contrainte d'être non corrélée aux précédentes. Les quatre coefficients utilisés pour calculer chaque composante principal constituent un *vecteur principal*. Il y a donc quatre vecteurs principaux. Ce sont les *vecteurs propres* de la matrice de corrélation. Leur calcul nécessite l'usage d'un ordinateur.

5) Compléter le tableau ci-dessous, calculé sur les valeurs centrées réduites des notes A, B, C et D :

	C_1	C_2	C_3	C_4
1	0.885	-1.706	0.269	0.000
2	-2.465	-0.119	0.696	0.000
3	-0.268	1.216	-0.260	-0.000
4	0.313	-0.222	-0.275	0.000
5	-2.449	-0.028	-0.422	-0.000

6	1.106	-0.472	2.116	0.000
7	0.253	2.249	0.387	-0.000
8	2.514	0.514	-0.756	-0.000
9	-0.208	-1.255	-0.921	0.000
10	0.321	-0.176	-0.834	0.000

Que peut-on dire de la 4e composante principale ? Calculer la moyenne et la variance des autres composantes principales, ainsi que leur coefficient de corrélation.

6) Représenter graphiquement dans un système d'axes orthonormé les articles caractérisés par les couples de valeurs (C_1, C_2). Que peut-on dire de la distance sur ce plan entre les articles 6 et 8 ? Ces axes sont appelés *axes principaux 1 et 2* et le plan *plan principal 1 x 2*. Calculer les distances sur ce plan entre les articles 5 et 8, entre 6 et 8 et entre 4 et 10. Calculer ces distances à l'aide des deux composantes principales C_1, C_2, puis des trois composantes principales C_1, C_2 et C_3.

Calculer la somme des carrés des distances entre les dix articles sur l'axe principal 1. Effectuer le calcul théorique. En déduire la somme des carrés des distances sur le plan principal 1 x 2. Que peut-on dire de l'axe principal 4 ?

7) On considère les deux variables C_1, A. Représenter les articles i = 1, …, 10 suivant les couples (C_1, A). Que peut-on en dire ? Même question pour C_2, B. Compléter le tableau des coefficients de corrélation entre les variables initiales et les composantes principales donnés ci-dessous :

	C_1	C_2	C_3	C_4
A	-0.796	-0.424	-0.431	0.000
B	-0.672	0.515	0.532	0.000
C	0.676	0.627	-0.388	0.000
D	0.732	-0.567	0.377	0.000

Calculer la somme des carrés des coefficients de corrélation des critères A, B, C et D avec la composante principale C_1. Que peut-on en dire ? Vérifier cette conjecture avec les autres composantes principales.

Calculer la somme des carrés des coefficients de corrélation du critère A avec les composantes principales C_1, C_2, C_3 et C_4. Que peut-on en dire ? Vérifier cette conjecture avec les autres composantes principales.

8) Représenter graphiquement dans un repère orthonormé les critères A, B, C et D en fonction de leurs coefficients de corrélation avec les composantes principales C_1 et C_2. On obtient ce que l'on appelle le *cercle de corrélation* $C_1 \times C_2$.