TD: ANALYSE EN COMPOSANTES PRINCIPALES

1. L'ACP permet:

□ A. D'étudier la variabilité des individus;

□ B. D'étudier les liaisons entre les variables

□ C. De regrouper les variables liées en nouvelles variables synthétiques pour réduire le nombre de colonnes de nos données

□ *D. Toutes les réponses*

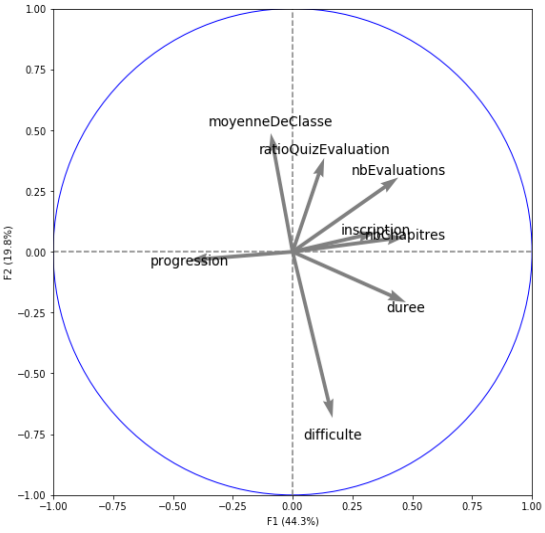
2. Parmi les affirmations suivantes, lesquelles sont recommandées pour l'algorithme d'analyse par composantes principales? Sélectionner la réponse correcte:

□A. L'ACP permet d'améliorer la visualisation des données puisqu'on réduit les données vers un espace à deux dimensions ou à 3 dimensions.

□B. L'ACP permet de réduire le risque de sur-apprentissage puisque on réduit le nombre de colonnes alors on améliore la performance du modèle.

□*C. Les réponses A et B sont correctes.*

□D. Aucune de ces réponses



3. Parmi les propositions suivantes laquelle est correcte:

A. La progression du cours est corrélée positivement avec la première composante.

B. La progression du cours est corrélée négativement avec la première composante.

C. La progression du cours est bien représentée par le plan factoriel.

D. Aucune de ces réponses.

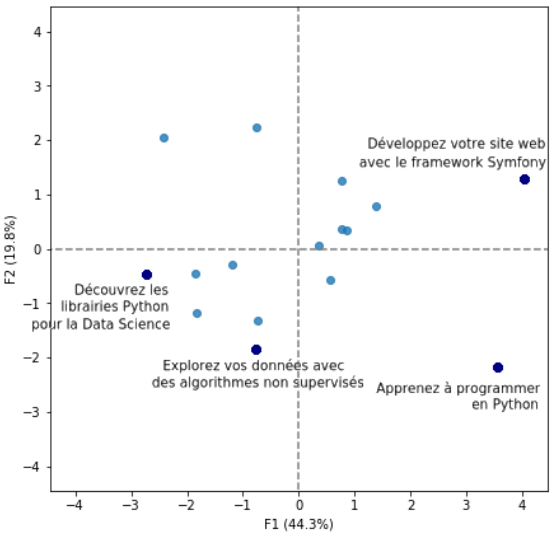
4. En superposant la carte des individus au cercle de corrélation, Quelle proposition est incorrecte **:**

A. *Les moyennes du cours "explorez vos données avec des algorithmes non supervisés" sont très bonnes.*

B. Le cours " Apprenez à programmer en Python" est très difficile.

C. Les moyennes des étudiants en "symphony" sont bonnes.

D. Le cours "apprenez à programmer en Python" est court.



5. Parmi les affirmations suivantes, lesquelles sont recommandées pour l'algorithme d'analyse par composantes principales? Sélectionner la réponse correcte :

A. L'ACP permet d'améliorer la visualisation des données puisqu'on réduit les données vers un espace à deux dimensions ou à 3 dimensions.

Wiem.trabelsi@esprit.tn

B. L'ACP permet de réduire le risque de sur-apprentissage puisqu'on réduit le nombre de colonnes alors on améliore la performance du modèle.

C. L'ACP permet de synthétiser les données représentées sur un échantillon de données.

D. *les. Les réponses A, B et C.*

E.

Exercice 1 :

Le tableau suivant fournit la structure du bilan d'un groupe pétrolier de 1969 à 1984 :

Année	NET	INT	SUB	LMT	DCT	IMM	EXP	VRD
1969	17.93	3.96	0.88	7.38	19.86	25.45	5.34	19.21
1970	16.21	3.93	0.94	9.82	19.11	26.58	5.01	18.40
1971	19.01	3.56	1.91	9.43	17.87	25.94	5.40	16.88
1972	18.05	3.33	1.73	9.72	18.83	26.05	5.08	17.21
1973	16.56	3.10	2.14	9.39	20.36	23.95	6.19	18.31
1974	13.09	2.64	2.44	8.10	25.05	19.48	11.61	17.59
1975	13.43	2.42	2.45	10.83	22.07	22.13	11.17	15.49
1976	9.83	2.46	1.79	11.81	24.10	22.39	11.31	16.30
1977	9.46	2.33	2.30	11.46	24.45	23.07	11.16	15.77
1978	10.93	2.95	2.25	10.72	23.16	24.17	9.64	16.20
1979	13.02	3.74	2.21	7.99	23.04	19.53	12.60	17.87
1980	13.43	3.60	2.29	7.09	23.59	17.61	16.67	15.72
1981	13.37	3.35	2.58	6.76	23.94	18.04	15.42	16.54
1982	11.75	2.74	3.11	7.37	25.04	18.11	14.71	17.18
1983	12.59	3.05	3.85	7.12	23.40	19.17	11.86	18.97
1984	13.00	3.00	4.00	7.00	24.00	20.00	12.00	17.00

Les postes de bilan sont les suivants :

NET : Situation nette ; représente l'ensemble des capitaux propres de l'entreprise.

INT : Intérêts ; représente l'ensemble des frais financiers supportés par l'entreprise.

SUB : Subventions ; représente le montant total des subventions accordées par l'Etat.

LMT : Dettes à long et moyen terme.

DCT : Dettes à court terme.

IMM : Immobilisations ; représente l'ensemble des terrains et du matériel de l'entreprise.

EXP : Valeurs d'exploitation.

VRD : Valeurs réalisables et disponibles ; ensemble des créances à court terme de l'entreprise.

Les données ont été ventilées en pourcentage par année, la somme des éléments d'une même ligne vaut 100, de manière à éviter les effets dus à l'inflation. On propose d'appliquer une Analyse en Composantes Principales (ACP) afin d'analyser l'évolution de la structure de bilan sur 15 ans. Les résultats de l'ACP sont présentés dans les tableaux et les figures ci-dessous :

Tableau1 :

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	4.47037150	2.35552573	0.5588	0.5588
2	2.11484576	1.43418677	0.2644	0.8232
3	0.68065899	0.17991239	0.0851	0.9082
4	0.50074660	0.34116829	0.0626	0.9708
5	0.15957831	0.09542998	0.0199	0.9908
6	0.06414833	0.05449844	0.0080	0.9988
7	0.00964990	0.00964928	0.0012	1.0000
8	0.00000062	0.0000	1.0000	

Tableau 3 :

Coordonnees des variables sur les axes

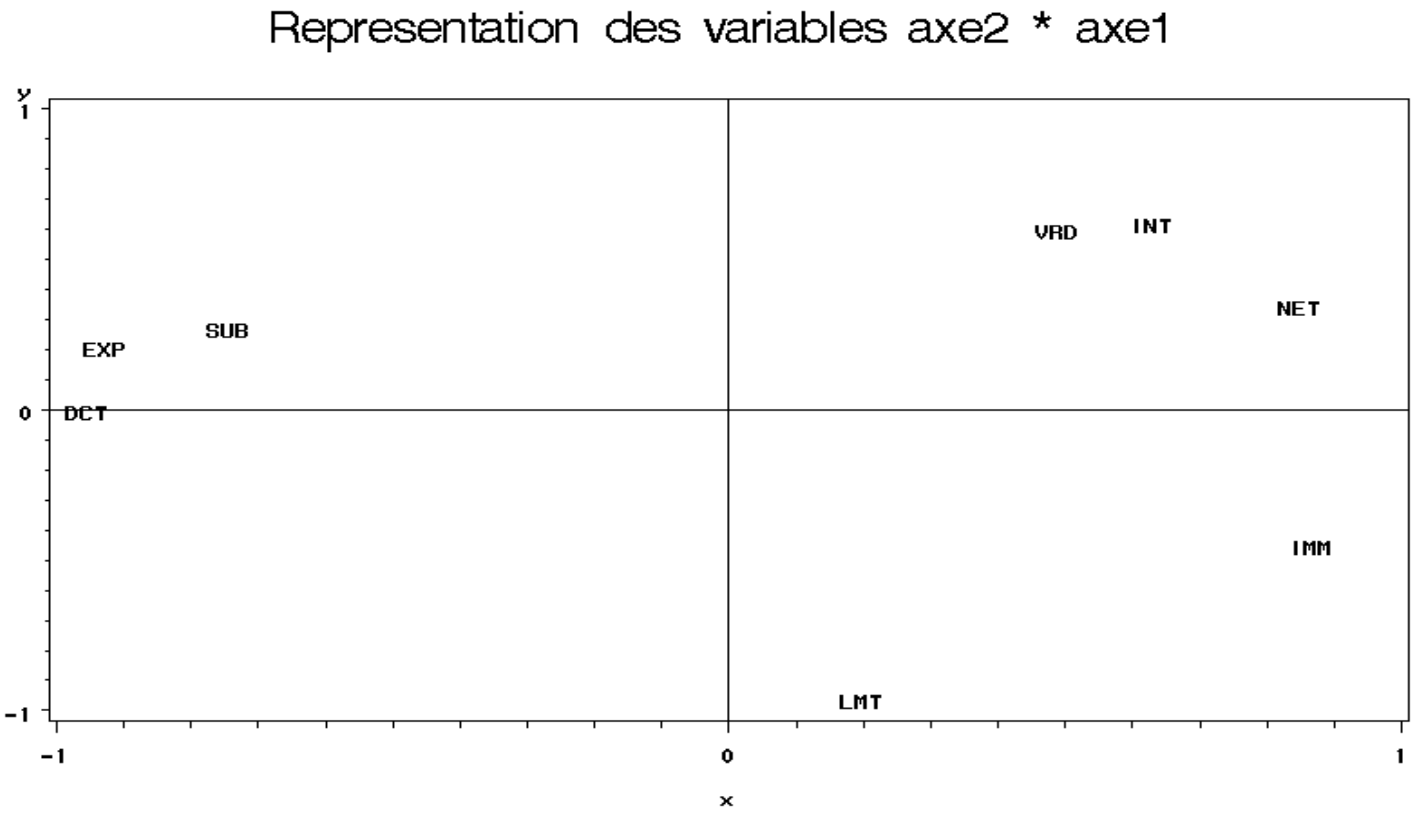
Pearson Correlation Coefficients, N = 16

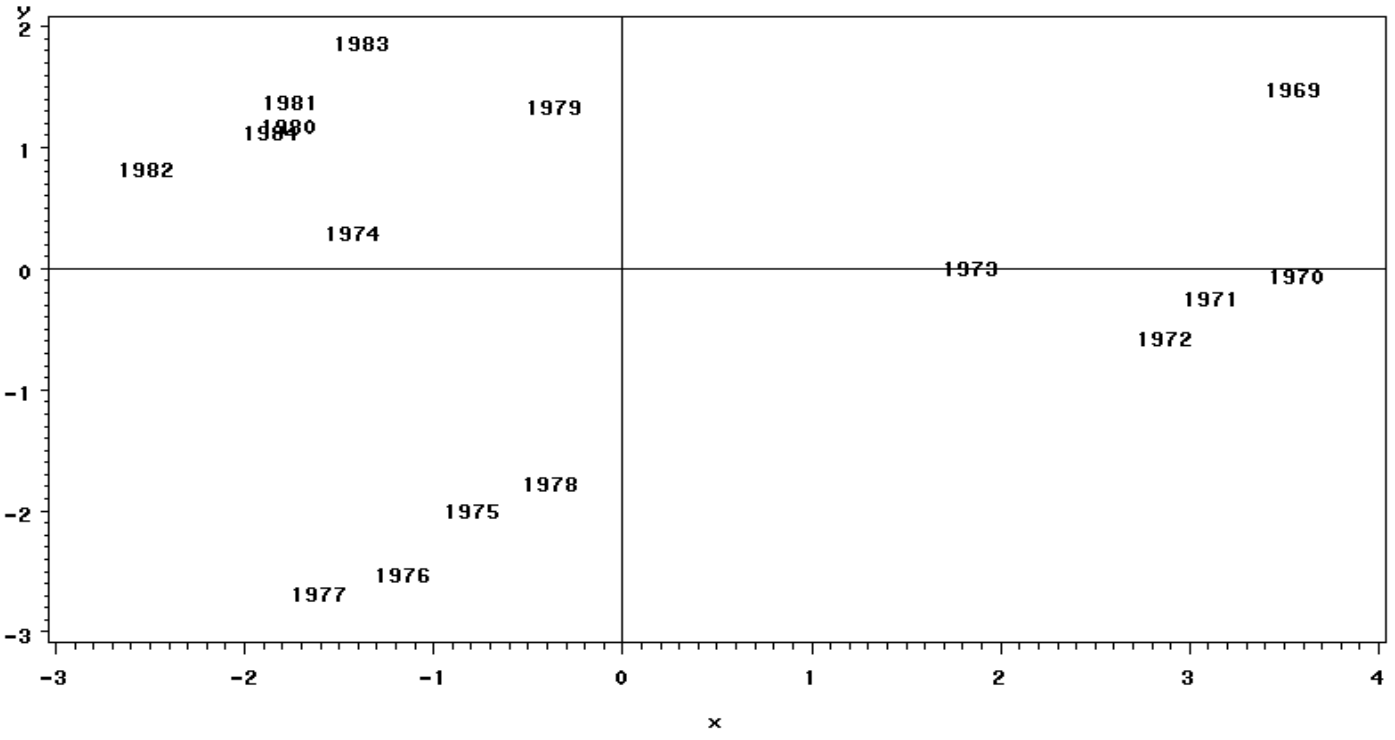
	Prin1	Prin2
NET	0.85014	0.34678
INT	0.62963	0.62173
SUB	-0.74214	0.27580
LMT	0.20017	-0.96163
DCT	-0.95386	0.00168
IMM	0.86787	-0.44767
EXP	-0.92571	0.20985
VRD	0.49025	0.60233

Tableau 2 :

Coordonnees et qualite de representation des individus sur les axes

annee	Prin1	Prin2	cos2_prin1	cos2_prin2
1969	3.55662	1.50535	0.78441	0.14052
1970	3.57546	-0.04273	0.93110	0.00013
1971	3.12027	-0.21808	0.83031	0.00406
1972	2.87553	-0.54758	0.89332	0.03239
1973	1.84936	0.02352	0.75517	0.00012
1974	-1.42432	0.32194	0.57269	0.02926
1975	-0.79476	-1.97215	0.11144	0.68621
1976	-1.16070	-2.50400	0.15851	0.73770
1977	-1.59726	-2.65758	0.25931	0.71786
1978	-0.37918	-1.74803	0.03739	0.79463
1979	-0.36150	1.35612	0.04004	0.56350
1980	-1.75965	1.20307	0.34868	0.16299
1981	-1.75001	1.40025	0.49152	0.31468
1982	-2.51840	0.84115	0.87166	0.09724
1983	-1.37918	1.88579	0.21797	0.40752
1984	-1.85228	1.15298	0.50000	0.19373





1. Expliquer les objectifs de l'Analyse en Composantes Principales (ACP) en Informatique décisionnelle.

2. Etude du tableau des valeurs propres.

2.1. A quoi correspond la somme des valeurs propres ?

2.2. On choisit de n'étudier que les deux premières composantes principales. Justifier ce choix en analysant le tableau 1 des valeurs propres (Eigenvalues).

2.3. Calculer le pourcentage d'information quantifié par les deux premières composantes principales sélectionnées.

3. Analyse des résultats de l'ACP

3.1. Sélectionner les individus (les années) qui sont bien représentés sur le plan factoriel en analysant les qualités de leurs représentations (cos2) dans le tableau 2.

3.2. Sélectionner les variables corrélées avec les premières composantes principales à partir du tableau 3.

3.3. Commenter les positions des années bien représentées sur le plan factoriel par rapport aux variables corrélées avec les deux premières composantes principales.

Exercice 2 :

On a rassemblé les résultats de 15 enfants de 10 ans à 6 subtests du WISC (scores 0 à 5). Les variables observées sont : CUB (Cubes de Kohs), PUZ (Assemblage d'objets), CAL (Calcul mental), MEM (Mémoire immédiate des chiffres), COM (Compréhension de phrases), VOC (Vocabulaire).

On traite ces données par une analyse en composantes principales normée. Les principaux résultats de cette ACP sont indiqués ci-dessous.

I. Etude du tableau des valeurs propres :

	Val. propr	% Total variance	Cumul Val. propr	Cumul %
1	3,2581	54,3020	3,2581	**54,3020**
2	1,8372	30,6194	5,0953	**84,9214**
3	0,4430	7,3831	5,5383	**92,3044**
4	0,2538	4,2292	5,7920	**96,5337**
5	0,1679	2,7990	5,9600	**99,3327**
6	0,0400	0,6673	6,0000	**100,0000**

Valeurs propres & statistiques associées

1. A quoi correspond la **somme** des valeurs propres ?

2. **<u>On choisit de n'étudier que les deux premières composantes principales</u>**. Justifier ce choix en analysant le tableau des valeurs propres.
3. **II.** Etude des qualités de représentation dans le premier plan principal

	Score Fact. 1	Score Fact. 2	Contribution Fact.1	Contribution Fact.2	Cos² Fact.1	Cos² Fact. 2
I1	-2,5616	3,0568	13,43	33,91	0,4078	0,5807
I2	-0,9661	0,9370	1,91	3,19	0,3907	0,3676
I3	0,6765	-0,6624	0,94	1,59	0,4446	0,4263
I4	-2,7969	-1,4636	16,01	7,77	0,7160	0,1961
I5	-1,8423	0,1211	6,95	0,05	0,8142	0,0035
I6	1,8891	0,1350	7,30	0,07	0,8426	0,0043
I7	-2,3396	-1,5487	11,20	8,70	0,6028	0,2641
I8	0,7275	-2,2054	1,08	17,65	0,0816	0,7499
I9	2,8400	0,5423	16,50	1,07	0,8745	0,0319
I10	2,1733	0,6117	9,66	1,36	0,7433	0,0589
I11	1,2940	2,0373	3,43	15,06	0,2256	0,5592
I12	-0,9947	0,8181	2,02	2,43	0,3120	0,2110
I13	-0,6099	-0,8730	0,76	2,77	0,1949	0,3994
I14	2,0150	-0,9470	8,31	3,25	0,7548	0,1667
I15	0,4957	-0,5591	0,50	1,13	0,1151	0,1464

Scores, contributions et qualités de représentation des individus

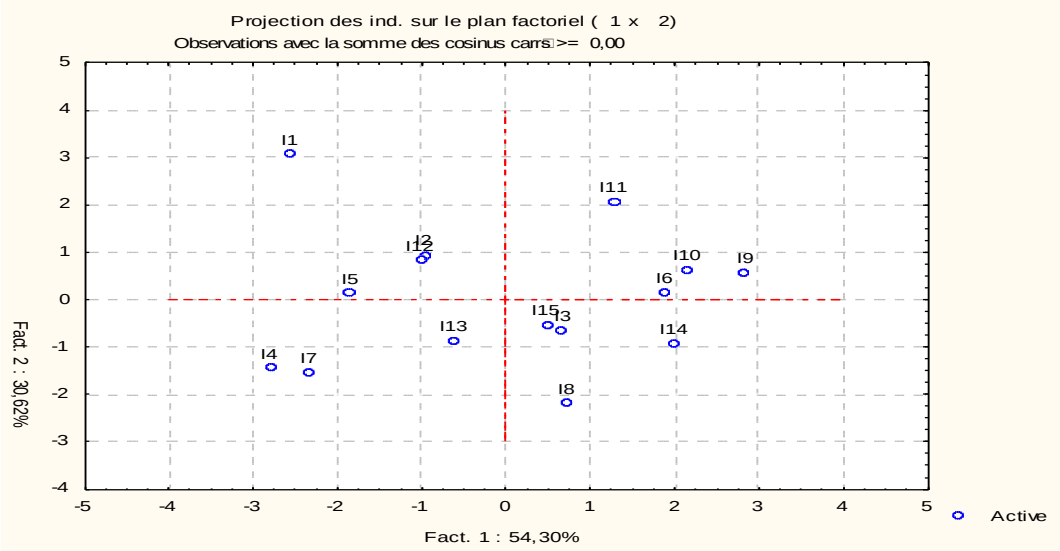
	Saturation Fact. 1	Saturation Fact. 2	Contribution Fact.1	Contribution Fact.2	Cos^2 Fact.1	Cos^2 Fact.1&2
CUB	-0,8970	0,2018	0,25	0,02	0,8046	0,8453
PUZ	-0,8652	0,2883	0,23	0,05	0,7485	0,8316
CAL	-0,9458	0,0390	0,27	0,00	0,8945	0,8960
MEM	0,4449	-0,7861	0,06	0,34	0,1980	0,8160
COM	-0,5382	-0,7627	0,09	0,32	0,2897	0,8714
VOC	-0,5683	-0,7156	0,10	0,28	0,3229	0,8350

Saturations, contributions et qualités de représentation des variables

3. Comment quantifie-t-on la qualité de représentation des individus par le plan factoriel ?

4. Quel est l'individu le moins représenté par le premier plan principal ? Quel est l'individu le mieux représenté ?

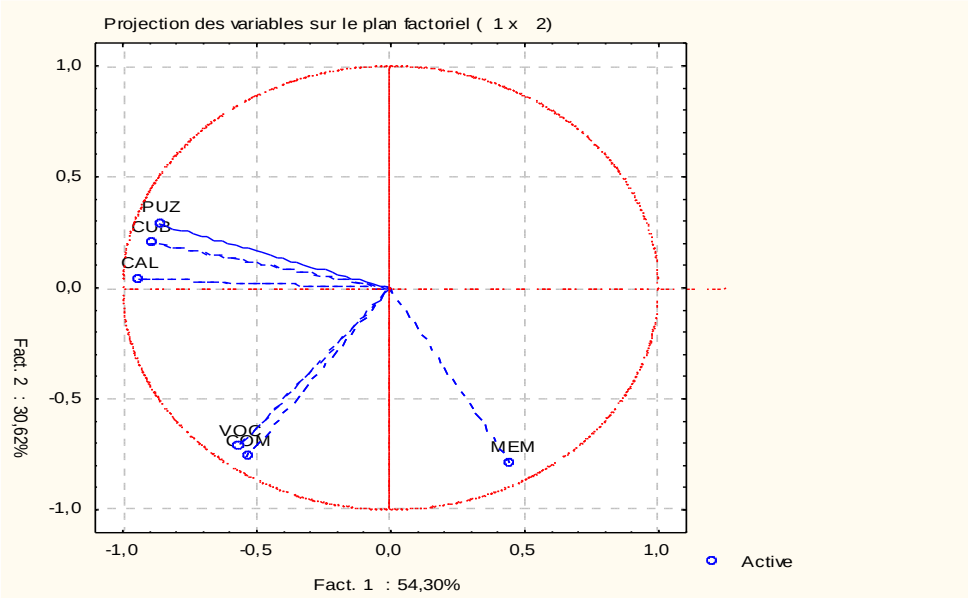
III. Etude du nuage des individus.



5. Quels sont les individus dont la contribution à la formation de la première composante principale est supérieure à la moyenne ? Pour chacun d'eux, préciser le signe de la coordonnée correspondante. Caractériser cet axe en termes d'opposition entre individus.

6. Même question pour la deuxième composante principale.

IV. Etude du nuage des variables



7. La représentation graphique des variables montre qu'elles sont toutes très bien représentées dans le plan (CP1, CP2). Justifier cette affirmation.

8. Quelles sont les variables qui sont corrélées positivement avec le premier facteur principal ? Quelles sont celles qui sont corrélées négativement ? Comment peut-on caractériser cet axe par rapport aux variables de départ ? (1.5)

9. Quelles sont les variables qui ont joué un rôle dominant dans la formation du deuxième axe

Wiem.trabelsi@esprit.tn