

Analyse en Composantes Principales EXERCICE DE SYNTHESE CORRIGE

Exercice 1



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). Le protocole observé est le suivant :

WISC	CUB	PUZ	CAL	MEM	COM	VOC
I1	5	5	4	0	1	1
I2	4	3	3	2	2	1
I3	2	1	2	3	2	2
I4	5	3	5	3	4	3
I5	4	4	3	2	3	2
I6	2	0	1	3	1	1
I7	3	3	4	2	4	4
I8	1	2	1	4	3	3
I9	0	1	0	3	1	0
I10	2	0	1	3	1	0
I11	1	2	1	1	0	1
I12	4	2	4	2	1	2
I13	3	2	3	3	2	3
I14	1	0	0	3	2	2
I15	2	1	1	2	3	2

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

Corrélations

	CUB	PUZ	CAL	MEM	COM	VOC
CUB	1,0000	0,7320	0,9207	-0,4491	0,3086	0,2735
PUZ	0,7320	1,0000	0,7510	-0,6143	0,2814	0,2850
CAL	0,9207	0,7510	1,0000	-0,3685	0,4077	0,4869
MEM	-0,4491	-0,6143	-0,3685	1,0000	0,3032	0,2023
COM	0,3086	0,2814	0,4077	0,3032	1,0000	0,7819
VOC	0,2735	0,2850	0,4869	0,2023	0,7819	1,0000

Val. Propres (matrice de corrél.) & stat. associées Variables actives seules

	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

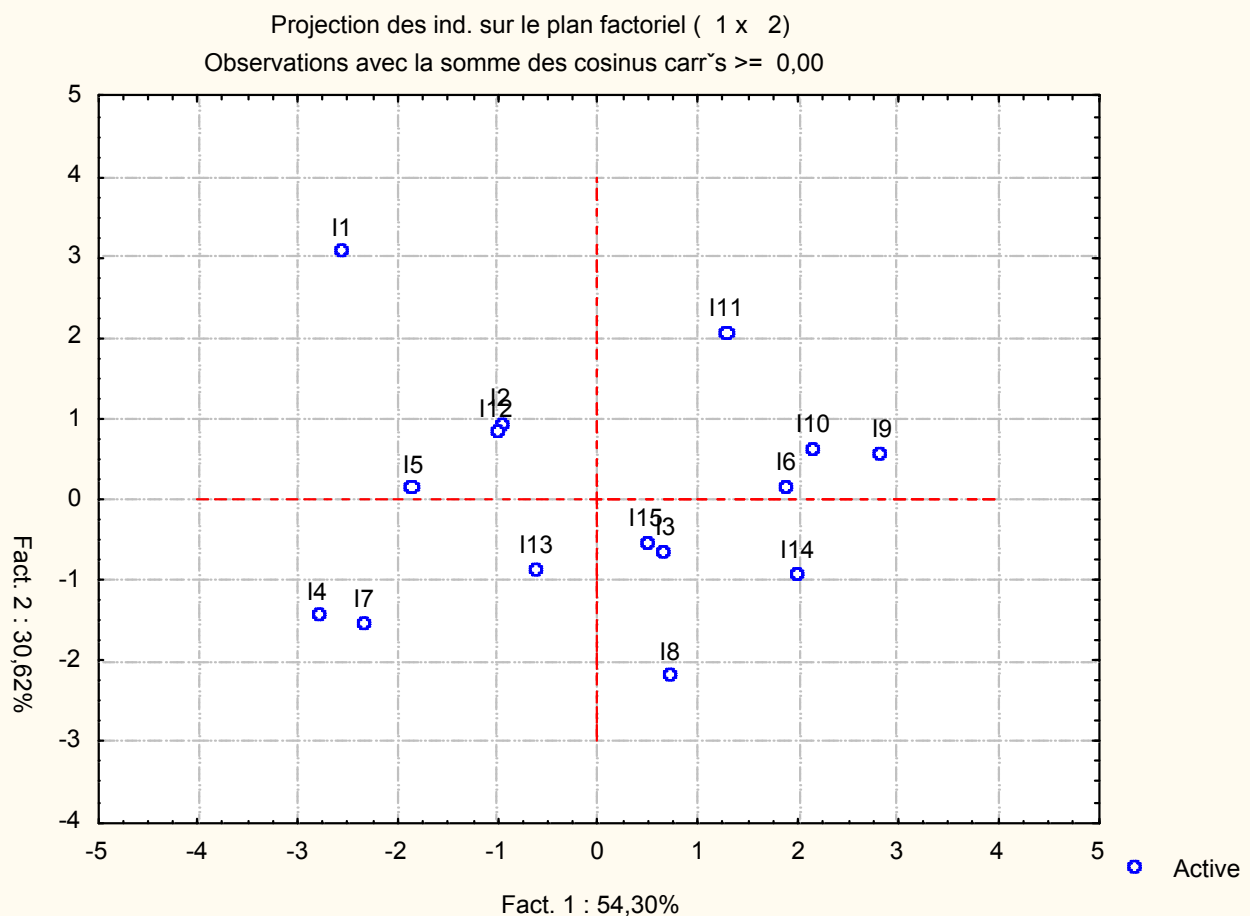
Scores, contributions et qualités de représentation des individus

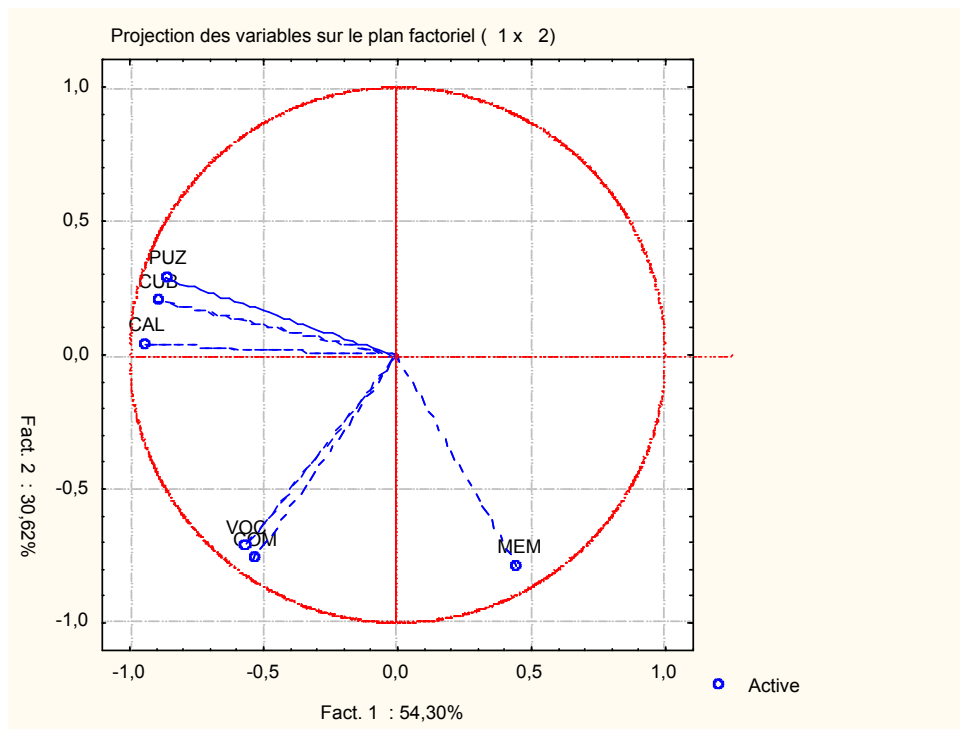
	Score Fact. 1	Score Fact. 2	Contribution Fact.1	Contribution Fact.2	Cos ² Fact.1	Cos ² Fact. 1&2
I1	-2,5616	3,0568	13,43	33,91	0,4078	0,5807
I2	-0,9661	0,9370	1,91	3,19	0,3907	0,3676
I3	0,6765	-0,6624	0,94	1,59	0,4446	0,4263
I4	-2,7969	-1,4636	16,01	7,77	0,7160	0,1961
I5	-1,8423	0,1211	6,95	0,05	0,8142	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	0,07499
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

Saturations, contributions et qualités de représentation des variables

	Saturation Fact. 1	Saturation Fact. 2	Contribution Fact.1	Contribution Fact.2	Cos ² Fact.1	Cos ² Fact.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

Représentation des individus dans le premier plan factoriel





1) Etude du tableau des valeurs propres

a) A quoi correspond la somme des valeurs propres ?

La somme des valeurs propres est égale à 6. Dans le cas d'une ACP normée telle que celle qui est effectuée ici, cette somme est égale au nombre de variables. Cette valeur correspond également à la variance du nuage des individus.

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

On peut choisir de ne s'intéresser qu'aux valeurs propres dont la contribution à la variance est supérieure à la moyenne. Ceci revient à étudier les composantes principales correspondant à des valeurs propres supérieures à 1. Or, seules les deux premières valeurs propres vérifient cette propriété sur l'exemple fourni.

2) Etude du tableau des corrélations

a) Quels sont les subtests les plus fortement corrélés entre eux ?

Les corrélations les plus fortes sont celles correspondant aux coefficients de corrélation les plus proches de 1 ou -1. Ici, le coefficient de corrélation dont la valeur absolue est la plus proche de 1 est celui qui relie CUB et CAL.

b) Comment s'organisent les signes "+" et les signes "-" dans le tableau des coefficients de corrélation. Commenter.

La plupart des coefficients de corrélation sont positifs. Autrement dit, un bon résultat à un test est, en règle générale, lié à de bons résultats aux autres tests. On note toutefois une exception remarquable : la variable MEM (mémoire immédiate des chiffres) est corrélée négativement à 3 autres variables : CUB, PUZ et CAL. Mais ce sont les seuls coefficients négatifs du tableau.

3) Etude des qualités de représentation dans le premier plan principal

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

Les qualités de représentation des individus dans le premier plan factoriel sont obtenues en additionnant les résultats des deux colonnes "Cos²" du tableau relatif aux individus. Le sujet le mieux représenté est I1, avec une qualité de représentation de 0,9885, le plus mal représenté est I15, avec une qualité de 0,2615.

4) Etude du nuage des individus.

a) 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.

Les individus envisagés ici sont ceux dont la contribution est supérieure à 6,67%. Ce sont les individus suivants :

-	+
I4 (16,01%)	I9 (16,5%)
I1 (13,43%)	I10 (9,66%)
I7 (11,2%)	I14 (8,31%)
I5 (6,95%)	

Caractériser cet axe en termes d'opposition entre individus.

Cet axe oppose les individus I4, I1, I7 et I5 d'une part aux individus I9, I10 et I14 d'autre part.

Les sujets du premier groupe obtiennent des scores généralement plus élevés que ceux du 2^e groupe, particulièrement sur les variables CUB, PUZ et CAL.

b) Même question pour la deuxième composante principale.

-	+
I8 (17,65%)	I1 (33,9%)
I4 (7,73%)	I11 (15,06%)

Cet axe oppose les individus I8 et I4 d'une part aux individus I1 et I11 d'autre part. On peut noter que l'individu I1 explique à lui seul plus du tiers de la variance de cet axe.

5) Etude du nuage des variables

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

Les extrémités des vecteurs représentant les variables sont toutes très proches du cercle des corrélations, ce qui montre que les qualités de représentation des variables sont satisfaisantes. D'autre part la dernière colonne du tableau des résultats relatifs aux variables montre que la variable la plus mal représentée est MEM, avec une qualité qui atteint quand même 0,8816.

b) 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 ?

Les corrélations peuvent être lues sur le dessin ou sur le tableau des saturations. Seule la variable MEM est corrélée positivement avec le premier axe, toutes les autres sont corrélées négativement. Ainsi, cet axe oppose le résultat au test MEM à tous les autres tests.

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

Les corrélations les plus fortes sont observées pour les variables MEM, COM et VOC. Ces variables sont toutes trois corrélées négativement avec la variable CP2.