

Correction ACP

Préparé par Sarra Zouari

QCM:

- 1D
- 2C
- 3B
- 4A

Exercice 1

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

L'objectif de l'ACP est la réduction de la dimensionnalité de n jusqu'à 2 ou 3 afin de trouver une nouvelle représentation / répartition des données pour en dégager les corrélations et les relations entre les individus et les variables.

2. Etude du tableau des valeurs propres.

2.1. A quoi correspond la somme des valeurs propres ? [0.5pt]

L'ensemble / totalité de l'information de l'entrepôt.

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

Selon le critère de Kaiser, on garde comme composantes principales celles qui ont des valeurs propres supérieures à 1.

2.3. Calculer le pourcentage d'information quantifié par les deux premières composantes principales sélectionnées. [0.5pt]

Le pourcentage de l'information qu'on garde avec les deux premières composantes principales est : 82.32% (cumulative)

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 (\cos^2) dans le tableau 2. [1pt]

Pour ce faire il faut calculer la moyenne des projections des variables sur le plan factoriel et puis garder ceux qui sont supérieur à la moyenne : la moyenne est 0.8

$\cos^2 \geq 0.8$: 1969, 1970, 1971, 1972, 1976, 1977, 1982

3.2. Sélectionner les variables corrélées avec les premières composantes principales à partir du tableau 3. [1pt]

Sur PRINT 1 : NET, INT, SUB, DCT, IMM, EXP

Sur PRINT 2: LMT, VRD

3.3. Commenter les positions des années bien représentées sur le plan factoriel par rapport aux variables corrélées avec les deux premières composantes principales. [1pt]

Les années (individus) 1969, 1970, 1971 et 1972 sont corrélées positivement avec NET, IMM et INT.

1982 est corrélée positivement avec EXP et DCT.

1976 et 1977 n'ont pas de corrélations.

Exercice 2

1/ L'ensemble / totalité de l'information de l'entrepôt. :

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.

2/ Selon le critère de Kaiser, ...

Etude du tableau des corrélations

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

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

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 général, 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/ par la somme des carrés des cosinus des coordonnées des individus sur les axes factoriels.

4 / Quel est l'individu le moins représenté par le plan factoriel? Quel est l'individu le mieux représenté ?

Les qualités de représentation des individus dans le 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.

5/ Etude du nuage des individus:

Quels sont les individus dont la contribution à la formation de la première composante principale est supérieure à la moyenne ?

Pour ce faire il faut calculer la moyenne des projections des variables sur la première dimension et puis garder ceux qui sont supérieur à la moyenne : la moyenne est 0.5

Les individus > à la moyenne plan 1 =0,54 (i4 i5 i6 i7 i9 i10 i14)

+ i6 i9 i10 i14

- i4 i5 i7

Les individus en bleu sont les mieux représentés

- 6/ Pour ce faire il faut calculer la moyenne des projections des variables sur la deuxième dimension et puis garder ceux qui sont supérieur à la moyenne : la moyenne est 0.3

Les individus > à la moyenne plan 2= 0,3 (i1 i2 i3 i8 i11 i13 i15)

+ i1 i2 i11

- i3 i8 i13 i15

Les individus en bleu sont les mieux représentés

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.

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.