LES CAHIERS DE L'ANALYSE DES DONNÉES

B. ESCOFIER

Traitement simultané de variables qualitatives et quantitatives en analyse factorielle

Les cahiers de l'analyse des données, tome 4, nº 2 (1979), p. 137-146

<http://www.numdam.org/item?id=CAD_1979__4_2_137_0>

Les Cahiers de l'Analyse des Données
Vol. IV - 1979 - n° 2 - p. 137-146

TRAITEMENT SIMULTANÉ DE VARIABLES QUALITATIVES ET QUANTITATIVES EN ANALYSE FACTORIELLE [QUALITATIVES ET QUANTITATIVES]

par B. Escofier (1)

1 *Résumé*

Quand on désire traiter simultanément des variables qualitatives et des variables quantitatives, la technique couramment utilisée est de coder les premières pour les transformer en variables qualitatives et de traiter l'ensemble par l'analyse des correspondances.

Nous rappelons d'abord les analogies et les différences entre le traitement de variables quantitatives par l'analyse en composantes principales et le traitement de ces mêmes variables par l'analyse des correspondances après codage.

Puis nous proposons une technique permettant de traiter simultanément les deux types de variables en conservant à chacune leur nature.

2 *Comparaison entre les deux analyses*

Soit un ensemble de p variables centrées et normées $V_1,\dots,V_p$ définies sur un ensemble I d'observations.

2.1 *En analyse en composantes principales*, chaque variable est représentée dans l'espace $\mathbb{R}^n$ par un vecteur. La métrique sur $\mathbb{R}^n$ est, au coefficient 1/n près, la métrique identité. Le cosinus de l'angle entre 2 vecteurs est égal à leur coefficient de corrélation.

Les facteurs sont les vecteurs de $\mathbb{R}^n$, orthogonaux deux à deux, maximisant l'expression :

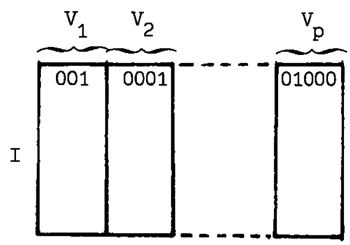
(1) B. ESCOFIER - I.N.S.A. - B.P. 14A - 35031 RENNES CEDEX - FRANCE

$$\sum_{s=1}^{p} \cos^2 \omega_s$$

où ω_s est l'angle entre le facteur et le vecteur V_1.

2.2 <u>*En analyse des correspondances*</u> on procède d'abord au <u>codage</u> de la variable, en divisant l'intervalle de $\mathbb{R}$ où la variable V_s prend ses valeurs en sous intervalle. Puis on indique pour chaque observation le sous intervalle dans laquelle elle se trouve. On obtient ainsi l'équivalent d'une variable qualitative à autant de modalités qu'il y a de sous intervalles.

Le tableau traité en analyse des correspondances croise les observations et les modalités des variables.



L'espace $\mathbb{R}^n$ est muni de la même métrique qu'en analyse en composantes principales puisque la marginale sur I du tableau disjonctif complet est constante.

Les variables n'apparaissent pas directement dans ce tableau, mais seulement à travers leurs modalités. Les colonnes associées à ces modalités sont égales aux variables indicatrices de la partition de I définie par la variable codée.

Une variable est donc représentée dans l'espace $\mathbb{R}^n$ par plusieurs points, en réalité le sous espace engendré par ces points.

Le premier facteur est le vecteur de $\mathbb{R}^n$ qui maximise l'expression :

$$\sum_{s=1}^{p} \cos^2 \alpha_s$$

où α_s est l'angle entre le facteur et le sous espace de $\mathbb{R}^n$ engendré par les modalités de la s-ième variable.

En effet, les modalités d'une même variable s'excluant entre elles, les vecteurs de $\mathbb{R}^n$ qui les représentent sont orthogonaux. L'inertie de chaque modalité est 1/p. Donc l'inertie de leurs projections dans une direction quelconque du sous espace qu'elles engendrent est constante et égale à 1/p. Et dans une direction quelconque de $\mathbb{R}^n$ elle sera égale au produit par 1/p du cosinus carré de l'angle entre cette direction et leur sous espace. Et, par définition, le premier facteur maximise l'inertie des projections des modalités de toutes les variables.

Remarque

Le sous espace engendré par les modalités de la variable V_s peut aussi être considéré comme le sous espace des fonctions de la variable codée $\tilde{V}_s$. On appelera variable codée $\tilde{V}_s$, toute variable constante sur chaque classe de la partition de I définie par le codage de V_s, cela peut être par exemple la moyenne de V_s sur chaque classe. En effet, une fonction de $\tilde{V}_s$ s'écrit comme la composée de $\tilde{V}_s$ par une fonction quelconque ψ de $\mathbb{R}$ dans $\mathbb{R}$: $\psi_o \tilde{V}_s$. Et l'ensemble de ces variables est exactement celui des variables constantes sur chaque classe de la partition.

Le premier facteur est le vecteur de $\mathbb{R}^n$ qui maximise la somme des cosinus carrés de ses angles avec les p sous espaces associés aux variables. C'est la variable la plus proche de toutes les variables $\tilde{V}_s$, non pas au sens d'une liaison linéaire comme en analyse en composantes principales, mais au sens d'une liaison fonctionnelle quelconque.

Avec les résultats de l'analyse des correspondances, on peut calculer très facilement les valeurs de $\cos^2\alpha_s$ qui expriment à quel point le facteur F est proche d'une fonction quelconque de $\tilde{V}_s$. En effet, $\cos^2\alpha_s$ est égal au quotient de l'inertie des projections des modalités de la variable V_s sur F par l'inertie totale de ces modalités. Ceci peut aider à interpréter les résultats et à mettre en évidence des liaisons non linéaires entre des variables.

La fonction de $\tilde{V}_s$ la plus proche du facteur F est donnée directement dans les résultats, c'est une variable sur I, constante sur chaque classe de $\tilde{V}_s$, c'est la valeur du facteur pour les modalités de $\tilde{V}_s$.

3 *Traitement simultané*

3.1 *Le principe*

Coder une variable quantitative d'une part, pose des problèmes de codage, choix de la partition, etc... ; d'autre part entraîne une perte d'information. Mais surtout, on a bien mis en évidence le fait que la variable numérique disparaît pour être remplacée par l'ensemble de toutes les fonctions de cette variable codée. Donc la seule information utilisée est la tribu engendrée sur I par cette variable, en réalité une sous tribu puisque la variable est éclatée en un petit nombre de modalités. Donc on considère cette variable, non pas pour ses valeurs numériques, mais pour la structure qu'elle définit sur l'ensemble des observations.

Dans certains cas, la nature de la variable numérique peut justifier ce point de vue, mais dans d'autres cas, le codage en variable qualitative n'est qu'une technique permettant de calculer des facteurs à partir de variables de différents types. Or, d'après ce qui précède, traiter une variable en variable quantitative, c'est lui associer un sous espace de $\mathbb{R}^n$ de dimension 1. La traiter en variable qualitative, c'est lui associer un sous espace de dimension supérieure à 1.

Ensuite, chercher les facteurs, c'est chercher les vecteurs de $\mathbb{R}^n$ les plus proches de l'ensemble des sous espaces considérés, c'est à dire les vecteurs qui maximisent la somme

$$\sum_{s=1}^{p} \cos^2 \omega_s$$

où ω_s est l'angle entre le vecteur et le sous espace de dimension 1 ou plus associé à la s-ième variable.

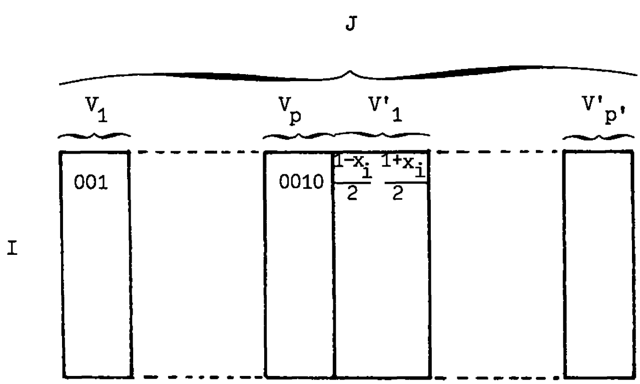
De ce point de vue, il n'y a aucune différence de nature entre les types de variables. On conçoit très bien de considérer simultanément des sous espaces de dimension 1 décrivant des variables quantitatives et des sous espaces de dimensions supérieures décrivant des variables qualitatives. Il suffit maintenant, le problème étant ainsi posé, de trouver une technique simple permettant les calculs des facteurs. Le programme classique d'analyse des correspondances le permet en codant judicieusement les variables.

3.2 *La technique*

Les variables considérées comme qualitatives sont codées comme toujours par des tableaux disjonctifs complets.

Une variable qualitative sera codée par deux colonnes. Dans l'une on mettra $\frac{1-x_i}{2}$ et dans l'autre $\frac{1+x_i}{2}$ où x_i est la valeur de la variable quantitative normée centrée.

Le tableau codant p variables qualitatives et p' variables quantitatives croise l'ensemble I des observations et un ensemble J qui comprend les modalités des p variables qualitatives et les couples associés aux p' variables quantitatives :



3.3 *Démonstration*

a) Les sommes de chacune des 2 colonnes associées à une variable quantitative sont égales à n/2. Il est nécessaire que ces sommes soient non nulles pour le programme d'analyse des correspondances.

b) Pour toute observation i de I, la somme $\frac{1-x_i}{2} + \frac{1+x_i}{2}$ est égale à 1. La marginale du tableau est donc égale au nombre total p+p' de variables. La métrique induite sur $\mathbb{R}^n$ dans l'analyse des correspondances de ce tableau est, à un coefficient près, la métrique identité. Le centre de gravité du nuage associé à l'ensemble J est situé sur la droite de vecteur directeur (1,1,...,1).

c) Notons X le vecteur de $\mathbb{R}^n$ représentant directement une variable quantitative normée et centrée. Notons j_1 et j_2 les 2 points de $\mathbb{R}^n$ représentant les profils des deux colonnes codant X dans notre tableau. On a les relations :

$$Oj_1 = OG-X$$

$$Oj_2 = OG+X$$

Sur tout vecteur orthogonal à OG, donc sur tous les facteurs de l'analyse, les projections de j_2 et de j_1 sont symétriques par rapport à G et confondues respectivement avec les projections de X et de son opposé.

L'inertie de j_1 ou de j_2 par rapport au centre de gravité du nuage est égale à 1/2(p+p'). L'inertie du couple vaut donc 1/(p+p'). L'inertie de sa projection dans une direction quelconque est égale au cosinus carré de l'angle entre X et cette direction multipliée par 1/(p+p').

Si ω_s note l'angle entre un facteur F et le sous espace engendré par les modalités de la s-ième variable qualitative et ω'_s l'angle entre F et la variable V'_s, l'inertie de F est :

$$\frac{1}{p+p'} \{ \sum_{s=1}^{p} \cos^2 \omega_s + \sum_{s=1}^{p'} \cos^2 \omega'_s \}$$

Comme le premier facteur sur I a une direction qui rend cette inertie maximum, nous obtenons bien les fonctions cherchées.

4 *Exemple*

4.1 *Les données*

C'est l'étude de plantes aquatiques, des typhas. Sur 43 de ces plantes, ont été relevées 10 mesures, 2 variables qualitatives et 3 nombres d'incidence. (voir le tableau des données brutes page suivante).

Nous les avons traitées en conservant aux 10 mesures leur nature de variables quantitatives, c'est à dire en les codant par les 2 colonnes $\frac{1-X_i}{2}$ et $\frac{1+X_i}{2}$. Nous avons éclaté en 5 classes, chacune des 3 variables correspondant aux nombres d'incidence . Les 2 variables qualitatives ont respectivement 2 et 3 modalités. Le tableau traité a donc 40 colonnes et représente 10 variables quantitatives et 5 qualitatives. Pour voir le comportement des 3 nombres d'incidence, quand on les considère comme des variables quantitatives, nous avons mis en éléments supplémentaires les 3 colonnes $\frac{1+X_i}{2}$ les concernant.

ETUDE DE TYPHAS DANS LA GRANDE MARE DE BEAULIEU

A hauteur totale

B longueur de l'épi ♀

C diamètre de l'épi ♀ en son milieu

D largeur du limbe de la 3ème feuille en son milieu

E épaisseur de cette feuille au même niveau

F nombre de cloisons dans cette feuille au même niveau

G nombre de fruits fertiles sur 50 éléments

H moyenne de la mesure de 30 fruits

I distance de la gaine de la 1ère feuille à la base de l'épi ♀

J diamètre de la tige à 20 cm sous l'épi ♀

K nombre de microns sur 30 carpodiens

L couleur de l'extrémité des poils

M intervalle entre l'épi ♂ et l'épi ♀

N forme du stigmate

O hauteur d'eau

Premières lignes du tableau des données brutes

	A cm	B cm	C cm	D mm	E mm	F	G	H mm	I cm	J mm	K	L Bbrun bblanc	M cm	N a:alène sp:spatule	O
1	195	14	1.8	3.5	1	9	21	4.03	19.5	3.9	11	b	4.1	a sp	39
2	151	20.5	1.35	5	1.5	9	26	3.1	0	7.5	12	b	4.5	a	48

Premières lignes du tableau des données codées. On n'indique que les premières colonnes, les variables qualitatives sont les premières. Pour les variables quantitatives le signe % indique la colonne $\frac{1-X_i}{2}$

	K					G					F					L		N			A %	A	B %	B
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	1	2	3				
																B	b	a	sp	a sp				
1	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0	0	1	0	0	1	0.222	0.778	0.901	0.099
2	0	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0	1	1	0	0	1.185	-0.185	0.301	0 699

4.2 *Les facteurs sur l'ensemble des variables*

Dans les résultats, les variables numériques sont représentées par 2 points correspondant respectivement aux colonnes $\frac{1+X_i}{2}$ et $\frac{1-X_i}{2}$. Les projections de ces deux points sur les facteurs sont confondus respectivement avec la projection de la variable X et de son opposé -X. Ils sont donc symétriques par rapport à l'origine et leurs coordonnées sur un facteur sont égales au coefficient de corrélation de X et du facteur affectés des signes + et - .

Dans le graphique du plan des 2 premiers facteurs donné à la page suivante, on a conservé seulement la colonne $\frac{1+X_i}{2}$, i.e. la variable X, et tracé une flèche joignant l'origine à ce point. Ceci permet de repérer les variables numériques. Ces variables sont notées A, B, C, D, E, H, I, J, M et O. Les 3 variables supplémentaires, les nombres d'incidence sont notés F, G, K. Nous avons tracé aussi le cercle de rayon 1.

Les variables qualitatives sont représentées exactement comme dans un tableau disjonctif complet par l'ensemble de leurs modalités. Ces modalités ont leur centre de gravité à l'origine. Nous avons ici 5 variables qualitatives notées L, N, F, G, K, les 3 dernières proviennent d'un codage par classe. Le chiffre suivant la lettre indique la modalité. Les classes de F, G, K sont prises dans l'ordre croissant.

Le groupe de variables C, D, J, H, et les deux variables supplémentaires K et F sont assez correlées négativement avec le premier facteur. Ce facteur oppose la modalité 5 de K à toutes les autres modalités de K. Et les modalités 4 et 5 de F aux autres modalités de F. Par contre, il ne respecte même pas l'ordre des modalités de G qui est d'ailleurs très peu corrélé avec ce premier facteur. Les modalités des 2 variables qualitatives sont dans le 1.2, situées pratiquement sur ce premier facteur auquel elles sont très liées. Ce facteur s'interprète donc à la fois comme une fonction très correlée au groupe de variables numériques cité ci-dessus, correlé négativement avec M, et comme un facteur opposant certaines modalités des variables qualitatives.

On remarque que F et K qui, en tant que variables numériques sont représentées par des points très proches se différencient nettement en tant que variables qualitatives : 2 modalités de F, F_4 et F_5, sont situées à gauche de l'origine, l'une F_5 est très loin du centre de gravité, alors qu'une seule modalité de K, K_5 est à gauche et proche de F_4.

Le codage en classe a donc mis en évidence pour ces deux variables une structure très différente. Nous avons ici mis en variables supplémentaires les variables numériques ; on peut aussi travailler sur les variables numériques et mettre en éléments supplémentaires les variables codées par classes, et même pourquoi pas, avec plusieurs codages différents pour étudier finement les liaisons entre variables.

Notons $\cos^2(K, \mathfrak{F}_1)$ le cosinus carré de l'angle entre la variable K et le facteur $\mathfrak{F}_1$, et $\cos^2(\tilde{K}, \mathfrak{F}_1)$ le cosinus carré de l'angle entre $\mathfrak{F}_1$ et le sous espace de dimension 4 associée à la variable codée K. Et comparons ces angles.

Le premier est donné dans le listage des résultats, le second, dans les programmes actuels nécessite un petit calcul.

Le tableau ci-dessous indique les valeurs des facteurs et des contributions données sur le listage des résultats du programme pour la variable K.

	Premier facteur	$\cos^2$	Contribution à l'inertie en millième	Deuxième facteur	$\cos^2$	Contribution à l'inertie en millième
K_1	735	41	8	-456	16	5
K_2	345	41	7	-503	87	22
K_3	390	52	9	396	54	14
K_4	147	5	1	188	8	2
K_5	-1146	398	68	105	3	1
K	-590	350	39	-185	35	6

$$\cos^2(\tilde{K}, \mathfrak{F}_1) = \frac{\text{Contribution absolue des modalités de K à } \mathfrak{F}_1}{\text{Inertie des moadalités de K}}$$

$$= \frac{93}{1000} \times 0.3 \times 15$$

$$= 0.418$$

Cette valeur est supérieure à celle de $\cos^2(K, \mathfrak{F}_1)$ qui vaut 0.350. Le facteur est donc plus près d'une certaine variable constante sur chaque élément de la partition induite par K, que de la variable K.

Pour le deuxième facteur $\mathfrak{F}_2$, on a :

$$\cos^2(K, \mathfrak{F}_2) = 0.035$$

et

$$\cos^2(\tilde{K}, \mathfrak{F}_2) = \frac{44}{1000} \times 0.2 \times 15 = 0.132 \ .$$

Ce deuxième facteur n'est donc pas correlé à K, mais il existe cependant une certaine liaison entre eux : $\mathfrak{F}_2$ oppose la classe moyenne K_3 aux petites classes K_1 et K_2.

Le deuxième facteur $\mathfrak{F}_2$ est correlé positivement aux variables E, B, O et A et négativement à la variable supplémentaire G. Il est lié aux 3 variables codées par classes, surtout la variable G, mais il n'est pas du tout lié aux deux variables qualitatives.

Regardons maintenant l'influence respective du groupe des 5 variables qualitatives et du groupe des 10 variables quantitatives sur chaque facteur. La suite des contributions à l'inertie de l'ensemble des 5 variables qualitatives sur les 5 premiers facteurs est :

$\mathfrak{F}_1$	$\mathfrak{F}_2$	$\mathfrak{F}_3$	$\mathfrak{F}_4$	$\mathfrak{F}_5$
0.46	0.23	0.52	0.78	0.76

Cette contribution est moyenne sur les facteurs $\mathfrak{F}_1$ et $\mathfrak{F}_3$, faible sur $\mathfrak{F}_2$ et devient importante sur $\mathfrak{F}_4$ et $\mathfrak{F}_5$. L'importance des deux dernières peut s'expliquer par le fait que les sous espaces associés aux variables codées étant de dimension 4, il faut un nombre de facteurs assez grand -au moins égal à 4- pour en extraire l'inertie.

4.3 Les facteurs sur l'ensemble des observations

Ce sont des fonctions qui résument "au mieux" l'ensemble des variables en rendant maximum la somme des carrés des corrélations avec les variables numériques et des carrés des rapports de corrélation avec les variables qualitatives.

La forme particulière du tableau de données simplifie les formules de transition vers ces facteurs. Dans ces formules, la somme des termes correspondant aux variables qualitatives donne le centre de gravité des modalités de l'observation i.

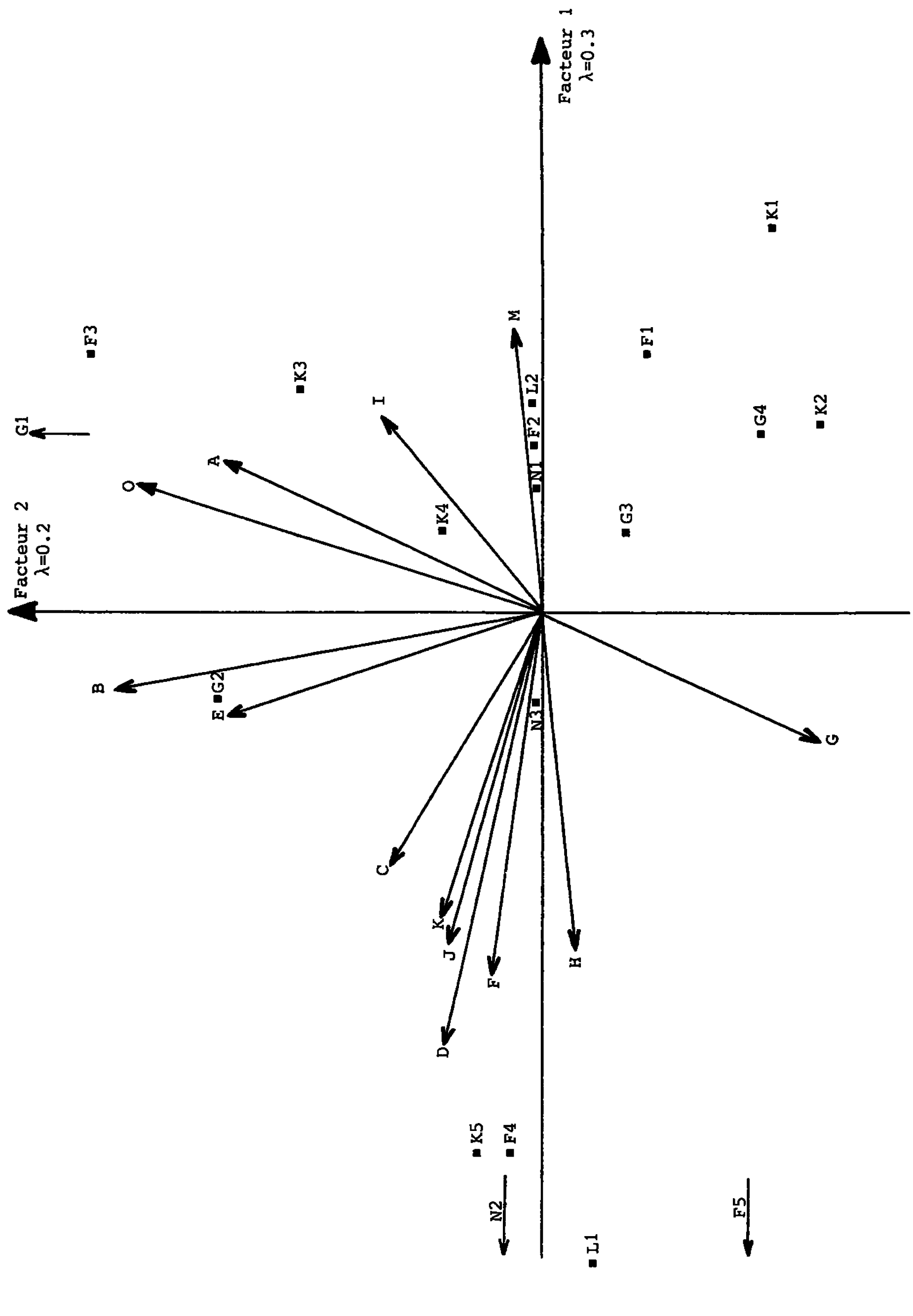
Notons X_i^j la valeur de la j-ème variable numérique normée centrée. Dans la formule de transition, la somme des deux termes correspondant à cette variable s'écrit :

$$\frac{1+X_i^j}{2}\,\mathfrak{G}(j) - \frac{1-X_i^j}{2}\,\mathfrak{G}(j) = X_i^j\,\mathfrak{G}(j)$$

5 Conclusion

Cette technique très simple qui utilise les programmes classiques paraît un moyen très efficace de traitement simultané de variables qualitatives et quantitatives. Ce problème n'avait pas jusqu'ici reçu de solutions satisfaisantes.

Quelques sorties supplémentaires des programmes pourraient aider l'interprétation des résultats : les corrélations entre les couples de variables numériques, les cosinus carrés des angles entre les facteurs et les sous espaces associés aux variables qualitatives.


Facteur 1
λ=0.3
Facteur 2
λ=0.2
A
B
C
D
E
F
G
H
I
J
K
M
O
F1
F2
F3
F4
F5
G1
G2
G3
G4
K1
K2
K3
K4
K5
L1
L2
N1
N2
N3