REVUE DE STATISTIQUE APPLIQUÉE

I. ANASTASSAKOS
G. D'AUBIGNY

L'utilisation des tests de sphéricité pour la recherche de la dimension de l'espace latent en analyse factorielle classique et en analyse en composantes principales

Revue de statistique appliquée, tome 32, n° 2 (1984), p. 45-57
<http://www.numdam.org/item?id=RSA_1984__32_2_45_0>

L'UTILISATION DES TESTS DE SPHERICITE POUR LA RECHERCHE DE LA DIMENSION DE L'ESPACE LATENT EN ANALYSE FACTORIELLE CLASSIQUE ET EN ANALYSE EN COMPOSANTES PRINCIPALES

I. ANASTASSAKOS(*), G. D'AUBIGNY(**)

(*) Laboratoire Statistique des Etudes Economiques et Sociales (LSEES). Commissariat à l'Energie Atomique. Fontenay-aux-Roses.

(**) Département Informatique et Mathématiques en Sciences Sociales. Université des Sciences Sociales de Grenoble.

RESUME

Un problème non résolu en analyse en composantes principales (ACP) ou en analyse factorielle classique (AF) est la recherche de la dimension de l'espace latent. Les simples règles empiriques (e.g. la règle de KAISER) et les tests statistiques (e.g. le test de BARTLETT) proposés, n'offrent pas de solutions satisfaisantes. Nous proposons ici une troisième approche, basée sur l'analyse du spectre des matrices de covariance. L'analyse de la variance des résidus, issus d'un modèle factoriel d'ordre K, obtenu par la minimisation des covariances partielles des variables observées sachant les k premiers facteurs, fournit une nouvelle interprétation de la statistique de VELICER, en termes de test de sphéricité des résidus. Les exemples d'application de cette statistique montrent qu'elle se comporte de deux façons différentes qui correspondent respectivement au concept de la structure simple de Thurstone et à celui de la description multidimensionnelle fournie par l'ACP conformément au modèle factoriel de Spearman.

I. INTRODUCTION

Divers tests statistiques ou de simples règles empiriques ont été proposés pour déterminer le nombre de facteurs qui engendrent le sous-espace latent, solution d'un problème de l'analyse factorielle.

Parmi les règles empiriques, fondées sur l'inspection directe du spectre des valeurs propres issues de l'analyse d'une matrice de corrélation, les plus souvent utilisées sont :

- la règle de KAISER [6] : sont retenus les facteurs associés aux valeurs propres supérieures à 1. Ce critère, très largement utilisé, trouve une justification partielle dans les travaux de L. GUTTMAN [5]. Cependant, GORSUCH [4] et FRANCISCO [3] opposent nombre de critiques à son emploi.
- le "scree test", proposé par CATTELL [2] : on examine le signe des écarts entre les différences de valeurs propres consécutives. Un changement de signe traduit, en effet, un point d'inflexion de la courbe des valeurs propres, indiquant ainsi le nombre de facteurs significatifs.

Mots Clés : **Dimension de l'espace latent, Tests de sphéricité, Analyse factorielle classique, Analyse en composantes principales, Analyse d'image.**

Notons $\{\lambda_j | j = 1, \ldots, p\}$ l'ensemble des valeurs propres de la matrice de corrélation des p variables observées. On a alors :

$$\begin{array}{llll} \delta_1 = \lambda_1 - \lambda_2 & & \epsilon_1 = \delta_1 - \delta_2 \\ \delta_2 = \lambda_2 - \lambda_3 & & \epsilon_2 = \delta_2 - \delta_3 \\ \vdots & & \vdots & \vdots \\ \delta_{p-1} = \lambda_{p-1} - \lambda_p & & \epsilon_{p-2} = \delta_{p-2} - \delta_{p-1} \end{array}$$

Le "scree test" retient m + 1 facteurs, m étant le plus petit élément de $\{1,2, \ldots, p-2\}$ pour lequel ϵ_m est négatif.

- L'examen de l'histogramme des valeurs propres (critère de coude) : dans cette variante graphique du "scree test" on retient les seuls facteurs dont le numéro d'ordre précède le coude où on remarque une moindre décroissance de l'inertie expliquée par les axes associés aux facteurs suivants.
- La part d'inertie expliquée : on se fixe un pourcentage minimum d'inertie que l'on veut restituer et on retient le nombre de facteurs en fonction de ce seuil.

La faiblesse majeure de ces méthodes résulte de leur caractère subjectif et la part d'arbitraire qu'impose leur mise en œuvre ; malgré leur simplicité elles se heurtent à un manque de fondements mathématiques ou statistiques. En particulier elles ne prennent en compte ni la taille de l'échantillon ni la méthode d'analyse choisie (analyse en composantes principales, analyse factorielle classique).

La méthode de validation croisée proposée par WOLD [11] se place, elle aussi, parmi les méthodes empiriques : on partage l'ensemble des observations en deux parties (échantillon de base et échantillon test) ; on effectue autant d'analyses factorielles sur la première partie qu'il y a de valeurs possibles du nombre de facteurs ; on applique les formules de régression ainsi obtenues sur la deuxième partie de l'échantillon. Le nombre de facteurs retenu correspond à la solution qui reconstruit au mieux les observations de cette deuxième partie. L'application de cette méthode est, cependant, assez lourde car elle necessite de nombreux calculs.

II. LES TESTS DE SPHERICITE

Les tests statistiques, construits pour déterminer la dimension de l'espace latent (BARTLETT, RIPPE, LAWLEY, JORESKOG), font appel pour la plupart d'entre eux à la notion de sphéricité de la matrice de variance-covariance des résidus et cela quels que soient l'objectif et la méthode d'analyse utilisée(*). Nous en exposons brièvement les principes :

Considérons une décomposition de la forme quadratique V associée à la matrice de variance-covariance des p variables observées $\{x_j, j = 1, \ldots, p\}$ fournie par une méthode d'analyse factorielle ; elle s'écrit :

$$V = V_1 + V_2 \qquad (1)$$

(*) Soit description comme en ACP soit modélisation comme en analyse factorielle classique ou en analyse d'image ou en ACP considérée comme approximation du modèle factoriel.

• V_1, de rang k, est la matrice de variance expliquée par les k facteurs

• $V_2 = V - V_1$ est la matrice de variance des résidus. Suivant le modèle choisi, V_2 est supposée diagonale (analyse factorielle classique) ou est de trace minimum (analyse en composantes principales).

Dans tous les cas, la sphéricité postulée de V, voire de V_2, conduit à tester successivement les deux hypothèses emboitées :

H_1 : la matrice de variance V est diagonale.

Lorsque H_1 est vérifiée et quelle que soit la méthode d'analyse utilisée on a :

$$V_2 = 0 \quad \text{et} \quad V = V_1$$

Dans le cas contraire une décomposition du type (1) plus élaborée est testée en formulant l'hypothèse :

H_2 : il existe une matrice V_1 de rang k fixé, telle que V_2 soit de la forme :

$$V_2 = \sigma^2 I, \text{ I étant la matrice identité.} \tag{2}$$

La généralisation immédiate de (2) suggère de retenir une matrice V_2 de la forme :

$$V_2 = D, \text{ D matrice diagonale.} \tag{3}$$

A toute décomposition de V respectant (3) est associée une décomposition respectant (2). Elle résulte de la transformation :

$$\left.\begin{aligned} V &= V_1 + V_2 \\ V_2 &= D \end{aligned}\right\} \Rightarrow \left\{\begin{aligned} &D^{-1/2}\, V\, D^{-1/2} = D^{-1/2}\, V_1\, D^{-1/2} + I \\ &D^{-1/2}\, V_2\, D^{-1/2} = I \end{aligned}\right.$$

La recherche d'une décomposition respectant (2) ou (3) est associée au test de l'hypothèse H_2 contre l'hypothèse alternative :

$$V_1 \text{ est de rang au moins } k + 1.$$

Si H_2 est rejetée, on reprend la procédure avec la valeur k + 1. La statistique de test q est le rapport des déterminants de V et de $\hat{V} = V_1 + \sigma^2 I$:

$$q = \frac{|V|}{|\hat{V}|} = \frac{\prod_{j>k} \mu_j}{\left(\frac{1}{p-k} \sum_{j>k} \mu_j\right)^{p-k}}, \quad \{\mu_j \mid j = 1, \ldots, p\} \text{ étant l'ensemble des valeurs propres de V.} \tag{4}$$

Une telle approche a été proposée en analyse en composantes principales (BARTLETT), en analyse factorielle d'image (JORESKOG) et en analyse factorielle classique (LAWLEY, RIPPE, RAO). Lorsque les variables $\{x_j\}$ suivent une loi de Laplace-Gauss p-dimensionnelle, la quantité :

$$2\ell nq \; \frac{N-1-k}{2}, \quad N \text{ étant la taille de l'échantillon,}$$

suit asymptotiquement une loi de χ^2 à $(p-k+2)(p-k+1)/2$ degrès de liberté sous l'hypothèse H_2 (KSHIRSAGAR [7]).

On dispose ainsi d'un moyen pour évaluer la significativité "statistique" des axes factoriels associés aux valeurs propres de V dans le cadre d'une décomposition du V du type (2).

Cependant, hormis les contraintes distributionnelles imposées pour valider cette famille de tests, la solution fournie est d'un intérêt limité (FRANCISCO [3]). Ces tests dépendent en effet de la taille N de l'échantillon de telle façon que pour N assez grand ils conduisent à un nombre excessif de facteurs significatifs.

La significativité "statistique" d'un axe factoriel ne doit pas être confondue avec son interprétabilité puisque elle fait référence à un modèle d'échantillonnage. Un axe significatif du point de vue statistique peut être sans importance pour la description du phénomène étudié au moyen d'une analyse factorielle. De plus une solution de (2) nécessitant un nombre de dimensions k peu inférieur au nombre de variables p est de peu d'intérêt pour le praticien.

III. ANALYSE EMPIRIQUE DE LA SPHERICITE

Une autre façon d'aborder le problème consiste à tester la validité d'une décomposition respectant (2) sans recours aux hypothèses probabilistes. Une statistique a été récemment proposée sans grandes justifications par VELICER [10]. L'objet de ce travail est d'en préciser le sens et d'étudier son comportement dans diverses situations.

Supposons que V_1 soit de rang k. Alors postuler la diagonalité de V_2, revient à imposer à chaque couple de variables x_j, x'_j, le respect de la condition :

$$\text{cov}\,(x_j, x'_j / c_1, c_2, \ldots, c_k) = 0 \qquad (5)$$

où $c_1, \ldots, c_k$ sont les facteurs associés aux k premières valeurs propres de V. Cette condition impose la nullité des covariances partielles des variables sachant les k facteurs, covariances qui ne sont rien d'autre que les éléments hors diagonaux de V_2.

On peut donc rechercher la diagonalité de V_2 en résolvant la fonction de minimisation :

$$\text{Min}\ \{\text{cov}^2\,(x_j, x'_j / c_1, c_2, \ldots, c_k)\} \qquad (5')$$

En considérant la matrice de corrélation résiduelle $R^{(k)}$ associée à V_2, on peut, pour résoudre (5'), adopter l'un des deux critères suivants :

1 – Rendre maximum le déterminant de $R^{(k)}$: $|R^{(k)}|$

2 – Minimiser la somme des carrés des éléments de $R^{(k)}$, c'est-à-dire minimiser $\text{tr}\,(R^{(k)2})$.

$|R^{(k)}|$ est maximum lorsque les corrélations résiduelles sont nulles, minimum lorsqu'elles sont égales à 1 ou – 1, sous la condition qu'aucun des résidus n'est combinaison linéaire des autres.

Inversement $\text{tr}\,(R^{(k)2})$ est minimum lorsque $R^{(k)} = I$, maximum lorsque les corrélations sont égales à 1 ou – 1.

Notons $\lambda_j^{(k)}$, $j = 1, \ldots, p$ les valeurs propres de $R^{(k)}$; le critère 2 équivaut alors à la minimisation de la variance des $\lambda_j^{(k)}$ en fonction de k. En effet, compte tenu de ce que $\bar{\lambda}^{(k)} = \frac{1}{p}\sum_j \lambda_j^{(k)} = \frac{1}{p}\text{tr}(R^{(k)}) = 1$, on a :

$$\min [\text{tr } R^{(k)2}] = \min \left[\sum_{i,j=1}^{p} r_{ij}^{(k)2} \right] = \min \sum_{j=1}^{p} \lambda_j^{(k)2}$$

$$\Leftrightarrow \min \left[\sum_j \lambda_j^{(k)2} - \sum_j' \lambda_j^{(k)} \right] = \min \left[\sum_{i \neq j} r_{ij}^{(k)2} \right]$$

Or $\frac{1}{p} \left[\sum_j \lambda_j^{(k)2} - \sum_j \lambda_j^{(k)} \right] = \frac{1}{p} \sum_j \lambda_j^{(k)2} - (\bar{\lambda}^{(k)})^2 = \Lambda^{(k)}$ est la variance empirique des valeurs propres observées. On a, par conséquent :

$$\min [\text{tr } R^{(k)2}] \Leftrightarrow \min [\Lambda^{(k)}] \qquad \text{c.q.f.d.}$$

On trouve dans [8] une analyse détaillée des propriétés de la quantité-critère associée à la trace de $R^{(k)2}$, en fonction de la nature des données (variables qualitatives ou quantitatives) et de la méthode d'analyse choisie (analyse en composantes principales ou analyse des correspondances).

La variance empirique des valeurs propres de la matrice des corrélations résiduelle atteint son minimum, en fonction du nombre k de facteurs retenus, lorsque les valeurs propres non nulles de $R^{(k)}$ sont toutes égales ; il y a alors sphéricité de V_2 au sens de (2).

Cette condition n'est presque jamais réalisée dans la pratique. Mais lorsque l'on utilise l'analyse en composantes principales ou l'analyse d'image comme approximation du modèle factoriel, il est raisonnable de supposer qu'il existe un k, petit par rapport à p, pour lequel cette condition est approximativement satisfaite.

Velicer propose de mesurer la validité de (5') à l'aide d'une moyenne des carrés des corrélations résiduelles, sachant k facteurs $\{c_j, j = 1, 2, \ldots, k\}$:

$$f^{(k)} = \frac{1}{p(p-1)} \sum_{i \neq j} r_{ij}^{(k)2} =$$

$$= \frac{1}{p-1} \left[\frac{1}{p} \sum_j \lambda_j^{(k)2} - (\bar{\lambda}^{(k)})^2 \right] = \frac{1}{p-1} \Lambda^{(k)} ; k = 1, \ldots, p$$

On choisit alors la valeur de k qui réalise le minimum de $f^{(k)}$.

Le calcul de VELICER est calqué sur le calcul de vecteurs propres par un algorithme de puissance itérée :

- A la 1ère étape on calcule :

$$E^{(1)} = R - c_1 \, c_1'$$

$$R^{(1)} = (\text{diag } E^{(1)})^{-1/2} \, E^{(1)} \, (\text{diag } E^{(1)})^{-1/2} \; ;$$

où $R^{(1)}$ est la matrice des corrélations partielles entre les p variables, sachant le premier facteur c_1. La quantité-critère $f^{(1)}$ s'écrit suivant la formule (6) avec $k = 1$.

- A la k-ième étape on calcule de façon identique :

$$E^{(k)} = E^{(k-1)} - c_k \, c_k'$$

$$R^{(k)} = (\text{diag } E^{(k)})^{-1/2} \; E^{(k)} \, (\text{diag } E^{(k)})^{-1/2} \; ;$$

la statistique $f^{(k)}$ est donnée par la formule (6).

La statistique $f^{(k)}$ n'est autre, à un coefficient près, que la variance des valeurs propres de la matrice de corrélation résiduelle $R^{(k)}$; son comportement en fonction de k est celui de la variance de p points disposés sur une droite. S'il existe a priori une infinité de comportements possibles, certains correspondent à des situations couramment rencontrées dans la pratique :

Cas 1 : Lorsque les k_0 premiers facteurs, issus d'une analyse en composantes principales ou d'une analyse d'image, sont associés à des groupes de variables faiblement corrélés d'un groupe à l'autre et fortement corrélés à l'intérieur des groupes, la variance $\Lambda^{(k)}$ augmente pour les valeurs de k comprises entre 1 et $k_0 - 1$ (cf. Fig. 1) ; après l'extraction de k_0 facteurs, la matrice de variance résiduelle n'est autre que celle composé par les spécificités (résidus) c'est-à-dire par des variables dont on peut supposer, sans perte de généralité, qu'elles sont non-corrélées (*)

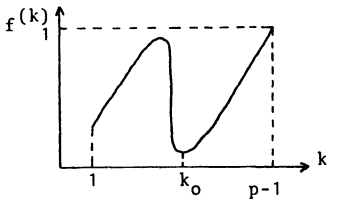


Figure 1

Le minimum de $f^{(k)}$ est alors atteint pour $k = k_0$; l'extraction des facteurs de rang supérieur à k_0 ($c_{k_0+1}, \ldots, c_{p-1}$), c'est-à-dire des axes associés aux résidus, provoque une nouvelle augmentation de la variance $\Lambda^{(k)}$. Par suite $f^{(k)}$ devient une fonction croissante de k, pour $k > k_0$.

Cas 2 : Les k_0 premiers facteurs de R ne sont pas associés à des variables structurées en groupes homogènes ; ils traduisent les effets principaux résultant des corrélations entre variables prises en compte de façon globale.

Dans ce cas (cf. Fig. 2) la variance $\Lambda^{(k)}$, $k = 1, \ldots, k_0$ décroît de façon continue puisque à chaque étape, de nombreuses variables sont assez bien corrélées

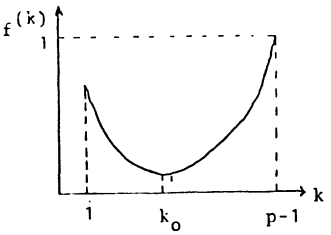


Figure 2

(*) Il s'agit là d'une situation identique à celle décrite par THURSTONE dans son modèle de structure simple [9].

avec l'axe correspondant ; pour des valeurs de k supérieures à k_0, on observe, comme dans le cas précédent une augmentation de la variance, due aux effets spécifiques (**).

Remarque : On observe parfois une oscillation autour du minimum, c'est-à-dire pour des valeurs de k au voisinage de k_0 (cf. Fig. 3). Il s'agit là d'un cas particulier du cas 2 pour lequel séparer effets principaux et effets spécifiques est difficile ; mieux vaut alors parler d'un intervalle $[k_1, k_2]$, que d'une valeur fixée k_0. La dimension retenue de l'espace latent sera choisie à l'intérieur de cet intervalle en fonction des caractéristiques propres de l'étude.

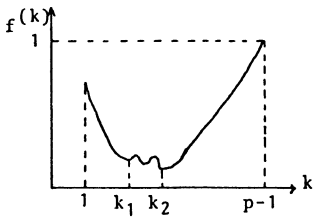


Figure 3

IV. APPLICATIONS

Le tableau I présente les résultats des tests effectués à partir de données qui diffèrent entre elles par la taille de l'échantillon et le nombre de variables traitées. La comparaison concerne l'analyse en composantes principales ainsi que l'analyse d'image.

On constate, à partir de ce tableau, que le test de VELICER fournit, en général, un nombre de facteurs inférieur à celui proposé par les autres tests. On remarque, en plus, la grande stabilité de ce test dans l'étude des données Seced ; de Seced 2 (66 variables, 1000 individus) à Seced 5 (80 variables, 2153 individus) le nombre de facteurs indiqué par le test est identique ($k_0 = 6$ facteurs). C'est dire que l'approximation du modèle ne gagne guère en qualité, lorsque l'on augmente le nombre de variables et la taille de l'échantillon.

Outre ces applications, l'efficacité du test proposé est jugée dans trois situations, servant de jauge pour évaluer ses possibilités.

Situation 1

Trois échantillons de taille N = 100 tirés au hasard dans une loi Gaussienne ($m = 0$, $\sigma = 1$) correspondent à trois variables gaussiennes indépendantes $\{x_i,$

(**) On rencontre ici le concept du facteur général, uni ou multidimensionnel, développé par SPEARMAN, dans son modèle d'analyse factorielle [9].

TABLEAU I

Les deux rubriques, "Risque" et "Seced", regroupent les données de deux enquêtes d'opinion publique réalisées au sein du LSEES ; dans la première on propose, aux interviewés, une liste de thèmes (situations diverses, maladies, jeux, etc.) et on leur demande d'attribuer à chaque thème, en fonction du risque qu'il présente, une note allant de 1 à 5 suivant une échelle d'accord en 5 paliers. Dans la deuxième enquête, répétée toutes les années depuis 1977 (Seced]), on fournit une liste de thèmes d'actualité sous forme de proposition et l'on demande aux individus d'exprimer leur sentiment sur chacun des thèmes, en lui attribuant une note sur une échelle du même genre ; cette enquête comporte un grand nombre de thèmes qui sont communs d'une année à l'autre.

	Analyse en composantes principales					Analyse d'image				
Données	Bartlett	Cattell	Kaiser	Coude	Velicer	Jöreskog	Cattell	Kaiser	Coude	Velicer
Risque 82 a N = 250, p = 40	38	5	11	5	7	26	4	23	4	7
Risque 82 b N = 250, p = 40	36	5	10	4	5-8	> 15	5	23	4	6-8
Risque 82 c N = 250, p = 40	37	5	9	5	7	> 15	2	23	5	7
Risque 82 d N = 250, p = 13	7	3	3	4	3	4	3	7	5	3
Risque 82 e N = 250, p = 13	11	5	4	5	2	4	3	7	4	2
Seced 1 N = 1000 p = 56	19	4	15	7	5	> 20	5	23	8	5
Seced 2 N = 1064 p = 66	27	5	18	5	6	> 20	5	29	7	6
Seced 3 N = 1622 p = 76	38	6	17	6	6	> 30	6	35	5	6
Seced 4 N = 528 p = 84	23	6	23	6	6	> 20	7	41	6	6
Seced 5 N = 2150 p = 80	52	8	19	7	6	> 30	8	39	7	6

$i = 1, 2, 3\}$. On simule de même 6 variables gaussiennes ($m = 0, \sigma = 0.01$) $\{u_j, j = 1, \ldots, 6\}$ indépendantes entre elles et indépendantes des x_i. Les x_i permettent de générer, à l'aide des coefficients de pondération a_{ij}, 6 variables y_j, combinaisons linéaires parfaites des x_i, auxquelles on ajoute les unicités correspondantes u_j. De façon formelle ceci s'écrit :

$$y_j = \sum_{i=1} a_{ij} x_i$$

$$y_j^* = y_j + u_j \quad ; \quad j = 1, \ldots, 6 .$$

Les coefficients a_{ij} et la matrice de corrélation des y_j^* sont donnés dans le tableau II. L'ensemble des y_j^* est soumis à une analyse en composantes principales sur matrice de corrélation. Chaque variable y_j^* est représentée presque parfaitement dans un même espace de dimension 3, à l'exception d'une petite partie qui ne représente que le onzième de sa variance et qui est dûe à la variable u_j qui lui correspond ; nous sommes ainsi placés dans la situation du premier cas présenté au paragraphe précédent.

TABLEAU II

Valeurs propres de R — 1

1	3.874
2	1.425
3	0.675
4	0.012
5	0.009
6	0.006

$f^{(0)} = 0.383009$

$\Lambda^{(0)} = 1.915045$

$E^{(1)}=R-c_1c_1'$; $S^{(1)}=(\text{diag } E^{(1)})^{-1/2}$

Valeurs propres de $S^{(1)}\ E^{(1)}\ S^{(1)} = R^{(1)}$ — 2

1	4.136
2	1.689
3	0.117
4	.031
5	0.025
6	0.000

$f^{(1)} = 0.466036$

$\Lambda^{(1)} = 2.33018$

$E^{(2)}=E^{(1)}-c_2c_2'$; $S^{(2)}=(\text{diag } E^{(2)})^{-1/2}$

Valeurs propres de $S^{(2)}\ E^{(2)}\ S^{(2)} = R^{(2)}$ — 3

1	5.299
2	0.344
3	0.246
4	0.110
5	0.000
6	0.000

$f^{(2)} = 0.742611$

$\Lambda^{(2)} = 3.713055$

$E^{(3)}=E^{(2)}-c_3c_3'$; $S^{(3)}=(\text{diag}E^{(3)})^{-1/2}$

Valeurs propres de $S^{(3)}\ E^{(3)}\ S^{(3)} = R^{(3)}$ — 4

2.760	
1.836	
1.403	
0.000	
0.000	
0.000	

$f^{(3)} = 0.232054$

$\Lambda^{(3)} = 1.16027$

$E^{(4)}=E^{(3)}-c_4c_4'$; $S^{(4)}=(\text{diag } E^{(4)})^{-1/2}$

Valeurs propres de $S^{(4)}\ E^{(4)}\ S^{(4)} = R^{(4)}$ — 5

3.981	
2.000	
0.024	
0.000	
0.000	
0.000	

$f^{(4)} = 0.461824$

$\Lambda^{(4)} = 2.30912$

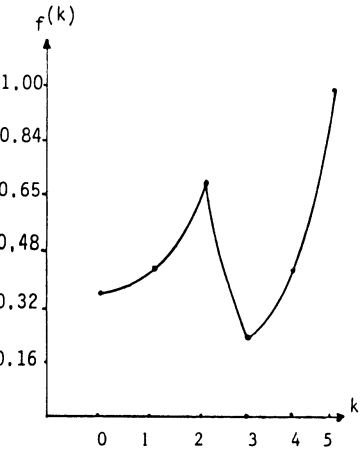
$E^{(5)}=E^{(4)}-c_5c_5'$; $S^{(5)}=(\text{diag } E^{(5)})^{-1/2}$

Valeurs propres de $S^{(5)}\ E^{(5)}\ S^{(5)} = R^{(5)}$ — 6

5.958	
0.045	
0.012	
0.000	
0.000	
0.000	

$f^{(5)} = 0.989489$

$\Lambda^{(5)} = 4.947445$



$y_1=.5x_1+.3x_2+.2x_3$

$y_2=.7x_1+.2x_2+.1x_3$

$y_3=.9x_1+0,8x_2+.02x_3$

$y_4=.3x_1+.1x_2+.6x_3$

$y_5=.01x_1+.8x_2+.19x_3$

$y_6=.2x_1+.35x_2+.45x_3$

Matrice des corrélations des y_i^*

1					
.928	1				
.821	.964	1			
.677	.532	.416	1		
.448	.150	.066	.215	1	
.752	.502	.309	.872	.645	1

Les résultats fournis par le test confirment notre analyse : la statistique $f^{(k)}$, $k = 0, \ldots, 5$ suit, en fonction de k, une courbe donnée par la figure 1 ; son examen suggère le choix de 3 axes, c'est-à-dire la dimension du sous-espace dans lequel se placent effectivement les variables y_j^*. Le pourcentage d'inertie expliquée par ces trois axes est de l'ordre de 99 %.

Situation 2

La deuxième application concerne le cas d'une situation d'équicorrélation :

$$R = \begin{bmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \vdots & & & \\ \rho & \rho & \ldots & 1 \end{bmatrix} \qquad ; \; 0 < \rho \leqslant 1$$

On sait (MORRISSON [9]) que l'analyse en composantes principales de cette matrice fournit une première valeur propre :

$$\lambda_1 = [1 + (p - 1)\rho]$$

tandis que les autres $p - 1$ valeurs propres sont toutes égales :

$$\lambda_2 = \ldots\ldots = \lambda_p = 1 - \rho$$

L'égalité des $p - 1$ dernières valeurs propres de R montre, d'après les tests de sphéricité, que l'on doit retenir un seul facteur, c'est-à-dire une approximation d'ordre 1 du modèle factoriel.

On a effectué l'analyse en composantes principales d'une matrice d'ordre 4 avec $\rho = 0.6$. Le test de VELICER suggère bien le choix d'un seul facteur ; la courbe d'évolution de la statistique $f^{(k)}$, en fonction du nombre k de facteurs est donnée par le tableau III.

Situation 3

Une autre matrice de corrélation possède une structure particulière, connue a priori. Elle correspond à la situation où les coefficients de corrélation multiple de chacune des p variables, sachant les $p - 1$ autres, sont tous égaux (BARGMANN [1]) ; les vecteurs propres d une matrice d'ordre p ainsi construite, fournissent les p contrastes orthogonaux du plan d'expérience à 2^p facteurs.

Un exemple de cette structure est donné par la matrice de corrélation

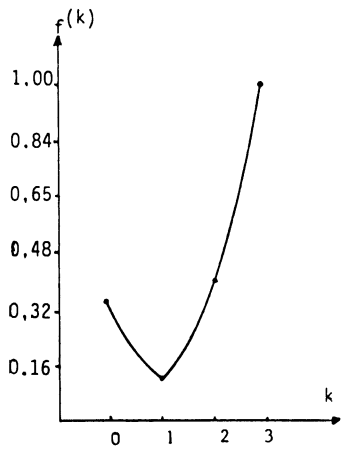
$$R = \begin{bmatrix} 1 & 0.7 & 0.6 & 0.4 \\ & 1 & 0.4 & 0.6 \\ & & 1 & 0.7 \\ & & & 1 \end{bmatrix}$$

dont les valeurs propres fournies par l'ACP sont :

$$\lambda_1 = 2.7$$
$$\lambda_2 = 0.7 \quad \lambda_3 = 0.5 \quad \lambda_4 = 0.1$$

TABLEAU III

k	Valeurs propres de $R^{(k)}$				$f^{(k)}$	$\Lambda^{(k)}$
	1	2	3	4		
0	2.8	0.4	0.4	0.4	0.36	1.08
1	1.33	1.33	1.33	0.00	0.11	0.33
2	2.67	1.33	0.00	0.00	0.41	1.22
3	4.00	0.00	0.00	0.00	1.00	3.00



Matrice de corrélation R

1.

0.6 1.

0.6 0.6 1.

0.6 0.6 0.6 1.

$f^{(k)}$ = quantité-critère de Velicer

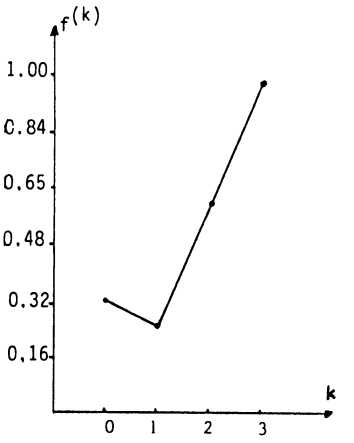
$\Lambda^{(k)}$ = variance des valeurs propres

$\Lambda^{(k)} = (p-1)\, f^{(k)}$; p = nb. de variables

c_1, c_2, c_3 = facteurs principaux de R

TABLEAU IV

k	Valeurs propres de $R^{(k)}$				$f^{(k)}$	$\Lambda^{(k)}$
	1	2	3	4		
0	2.7	0.7	0.5	0.1	0.33	1.01
1	2.15	1.54	0.31	0.00	0.26	0.78
2	3.34	0.67	0.00	0.00	0.63	1.69
3	4.00	0.00	0.00	0.00	1.00	3.00



Matrice de corrélation R

1.

0.7 1.

0.6 0.4 1.

0.4 0.6 0.7 1.

$f^{(k)}$=quantité-critère de Velicer

$\Lambda^{(k)}$ = variance des valeurs propres

$\Lambda^{(k)} = (p-1)\, f^{(k)}$; p = nb. de variables

c_1, c_2, c_3 = facteurs principaux de R

La grandeur relative des valeurs propres indique qu'il est raisonnable de postuler l'égalité des trois dernières ; le test de VELICER effectué à partir de R suggère le choix d'un seul facteur vérifiant ainsi l'hypothèse d'approximation d'ordre 1.

CONCLUSION

Le test de VELICER, considéré ici comme test de sphéricité de la matrice de variance résiduelle, issue d'une analyse en composantes principales ou d'une analyse d'image, offre une solution satisfaisante au problème du nombre de facteurs à retenir en analyse factorielle ; contrairement aux autres tests proposés, il est facile à calculer et fournit une interprétation claire du choix effectué, grâce à l'équivalence que nous avons établie entre la minimisation de la trace de $R^{(k)2}$ et la minimisation de la variance des valeurs propres de $R^{(k)}$ en fonction de $k (k = 1, \ldots, p - 1)$.

En analyse factorielle classique, les résultats du test peuvent être utilisés comme un point de départ pour l'ajustement de la matrice de corrélation R.

Nous discuterons, enfin, dans un prochain article, le cas d'application de ce test à l'étude de la liaison entre vecteurs aléatoires.

REFERENCES

[1] R. BARGMANN (1957). – *A study of independance and dependance in multivariate normal analysis*, University of North Carolina, Institut of Statistics Mimeographed Series n° 186, Chapel Hill, N.C.

[2] R.B. CATTELL (1966). – The scree test for the number of factors, *Multivariate Behavioral Research*, 1, 245-276.

[3] FRANCISCO C.A. and FINCH M.D. (1980). – A comparaison of methods used for determining the number of factors to retain in factor analysis, *Technometrics*, 105-110.

[4] R.L. GORSUCH (1973). – Using Bartlett's significance test to determine the number of factors to extract, *Educational and Psychological Measurement*, 33, 361-364.

[5] L. GUTTMAN (1954). – Some necessary conditions for common factor analysis, *Psychometrica*, 19, 149-161.

[6] H.F KAISER (1961). – A note on Guttman's lower bound for the number of common factors, *British Journal of Statistical Psychology*, 14, 1-2.

[7] A. KSHIRSAGAR (1972). – Likelihood ratio tests ; optimality of principal components, *in: Multivariate analysis*, Marcel Dekker Inc. New-York.

[8] A. LECLERC, P. AIACH (1978). – Mesure de l'importance des valeurs propres en analyse des données. Application à l'analyse en composantes principales de variables catégorisées, *Revue de Statistique Appliquée*, vol. XXVI, n° 1, 5-21.

[9] A. MORRISSON (1976). – The structure of multivariate observations, *in: Multivariate statistical methods*, New-York : McGraw-Hill.

[10] W.F. VELICER (1976). – Determining the number of components from the matrix of partial correlations, *Psychometrica*, 41, 321-326.

[11] S. WOLD (1978). – Cross validatory estimation of the number of components in factor and principal components analysis, *Technometrics*, 20, 397-405.