Indicateur informationnel synergique (*SYNAPSe*) pour tester la qualité et la signification de l'analyse factorielle sans perte d'information

Sabina Popescu – Spineni
Institut de Santé Publique & Université de Médicine
et Pharmacie "Carol Davila" - Bucarest, Roumanie
str. Dr. Leonte A., no. 1-3, sector 5, Bucarest
spopescu@ispb.ro

Résumé

Le but de l'analyse factorielle est de réduire les dimensions de la représentation spatiale (duale) des observations, en réduisant le nombre des axes factoriels. C'est très difficile d'établir un bon critère d'évaluation de la qualité d'une représentation dans l'espace réduit, après avoir réduit le nombre des axes. Le concept de *distance* joue un rôle essentiel dans la plupart des modèles d'analyse multivariée, en considérant les différentes directions de dispersion de la variabilité, mais la théorie des *valeurs propres* devient la plus importante dans l'analyse globale. En réalité, on applique des critères empiriques (Kaiser, Cattel), étant le cas, par exemple, de l'analyse des composantes principales pour les valeurs propres supérieures à 1, mais cette étude veut présenter une méthode pour déterminer et utiliser un indicateur informationnel agrégé, *SYNAPSe*[1] (Popescu-Spineni, 1998), qui a été construit par l'aide d'un concept de la statistique informationnelle, sans perte d'information, basé sur des références théoriques appropriées.

1 Introduction

Dans les sciences sociales, ainsi que dans l'économie ou dans le management, l'analyse multivariée des données a été imposée par des besoins opérationnels. Mais l'analyse des données utilise beaucoup de méthodes d'optimisation qui proposent des algorithmes très rigoureux pour établir la partition d'un set *n* d'objets caractérisés par *k* variables, qui caractérise une population ou un group de référence. L'analyse multidimensionnelle des données (MDA) peut inclure deux principales méthodes d'analyse: l'analyse cluster (AC) et l'analyse factorielle linéaire des données (composantes principales, canonique, analyse des correspondances), simples ou multiples. Dans ces méthodes on se propose une représentation synthétique avec le minimum de perte d'information. (Popescu-Spineni, 2000)

Le but de l'analyse factorielle, comme technique d'analyse multivariée, est de déterminer si la corrélation d'un grand nombre de variables observées peut être expliquée par un petit nombre de facteurs fondamentaux et combien de tels facteurs sont nécessaires. Des méthodes classiques d'optimisation sont adaptées pour suivre le type des données. (Benzécri, 1980)

Les solutions des méthodes mentionnées sont appropriées pour déterminer un nombre minimal de facteurs capables d'expliciter toute la variabilité, par une symétrie parfaite entre les profiles-colonnes et les profiles-lignes analysées, en considérant les contraintes et en envisageant les priorités. Le concept de *distance* joue un rôle essentiel dans la plupart des modèles d'analyse multivariée, en considérant les différentes directions de dispersion de la variabilité, mais la théorie des *valeurs propres* (eigenvalues) devient la plus importante dans l'analyse globale.

Le principal but de l'analyse factorielle est de réduire les dimensions de la représentation (duale) spatiale des observations, en réduisant le nombre des axes factoriels. Mais il est très difficile d'établir un critère adéquat pour mesurer la qualité de la représentation dans l'espace factoriel réduit, après avoir réduit le nombre des axes factoriels. (Lebart, 1995); (Rizzi, 1995)
Il y a beaucoup de techniques empiriques pour réduire le nombre des axes factoriels (Kaiser, Cattel, etc.), mais dans cet étude l'auteur présente un nouvel critère pour trouver et tester un indicateur informationnel agrégé (Popescu-Spineni, 1998), celui-ci calculé à l'aide d'un concept de la statistique informationnelle, sans avoir de perte d'information. De même, beaucoup d'exemples sont proposés pour appliquer et tester cet indicateur informationnel agrégé ou synergique (nommé *SYNAPSe*), avec des références théoriques appropriées.

2 L'analyse factorielle: la réduction du nombre des facteurs

La réduction de la dimension de l'espace (cartésien) des observations dans l'analyse factorielle ou des composantes principales revient à la réduction du nombre des axes factoriels, sans une solution rigoureuse, jusqu'à présent. (Lebart et collab., 1995)
Le but de l'analyse étant d'obtenir une représentation des observations dans l'espace de dimension réduite (soit < p), on doit apprécier la perte d'information après avoir calculé le nombre de facteurs retenus. On observe dans ce cas que le nuage des points n'est pas encore centré dans le centre de gravité (G). Après avoir calculé les valeurs propres (λ_i, avec $1 \leq i \leq p$) du système de k équations, on considère le critère du pourcentage de l'inertie totale (τ) comme mesure d'évaluation la qualité de la représentation factorielle (Benzécri, 1979), par :

$$\tau = \frac{\lambda_1 + \lambda_2 + ... + \lambda_p}{\lambda_1 + \lambda_2 + ... + \lambda_K} = \frac{\lambda_1 + \lambda_2 + ... + \lambda_p}{\varphi^2}.$$

Par exemple, avec $(\lambda_1 + \lambda_2)/(\varphi^2) = 0{,}9$, donc 90%, on voit que le nuage des points est aplatis sur un sous-espace à deux dimensions (Saporta, 1996). On peut observer que, dans l'appréciation du pourcentage de l'inertie, on doit tenir compte du nombre de variables initiales: un pourcentage de 10% n'a pas la même signification pour un tableau de 20 variables que pour un autre de 100 variables. (Benzécri, 1980)

Des critères théoriques ont été élaborés qui proposent de déterminer si les valeurs propres sont significativement différentes à partir d'un certain rang, sinon, on retient les premières valeurs propres (λ_i). La réduction de dimension n'est pas possible que s'il existe une redondance entre les variables $x_1, \ldots, x_p$, ou si elles sont indépendantes (Saporta, 1996). Par la méthode des approximations successives, on estime les valeurs propres et les vecteurs propres associés à

l'équation caractéristique $|R - \lambda I| = 0$, qui explique la plupart de la variation totale. En supposant que le processus s'est interrompu après avoir estimé p (p < K) valeurs propres, on doit vérifier si les plus petites des valeurs propres sont égales, afin de maximiser la variance des composantes principales de l'analyse. (Benzécri, 1980) En réalité, on applique des critères empiriques (Kaiser, 1965), étant le cas par exemple, de l'analyse des composantes principales pour les valeurs propres supérieures à 1.

Pour résoudre l'équation statistique matricéale Y = AX +ε. dans l'analyse factorielle, on propose un critère d'optimisation et la méthode d'estimation des paramètres. Pour l'estimation optimale des paramètres, il existe deux conditions principales: la reproduction des corrélations observées et l'explication de la variabilité. Les méthodes statistiques pour l'estimation des paramètres dépendent de la technique de résolution du système d'équations linéaires: (1) la méthode des moindres carrés; (2) le principe de la probabilité maximale (maximum likelyhood); - les deux étant équivalentes dans le cas normal (Gauss-Laplace). (Torrens-Ibern, 1972)

En général, dans l'analyse factorielle, le problème de tester le nombre de facteurs (ou d'axes) qu'on doit retenir se pose de deux directions: soit on suppose l'existence d'un nombre p de facteurs communs en vérifiant l'hypothèse nulle H_0, soit on détermine le nombre intégral de facteurs, en calculant leurs nombre à partir du nombre d'observations. (Lebart et collab., 1995)

Les plus importants critères théoriques proviennent de Barttlet (1951) et Lawley (1940), mais Jöreskog (1963) propose une procédure spécifique pour réduire le nombre de facteurs dans le cas des observations non-normales ("sans hypothèses préalables"). (Benzécri, 1980)

3 L'approximation de la distribution des valeurs propres

La réduction du nombre de facteurs dans l'analyse factorielle ou dans l'analyse des correspondances (Benzécri, 1980) dépend de l'analyse de la distribution théorique des valeurs propres (λ_i). Dans un tableau de contingence, sous l'hypothèse de l'indépendance des lignes et des colonnes, cette distribution s'approche d'une loi de distribution connue d'une matrice Wishart (Lebart, Morineau, Piron, 1995). On souligne que la loi de distribution des valeurs propres résultées de l'analyse a été appréciée de manière erronée, ainsi que l'inertie totale du nuage des points, comme de type χ^2, tandis que différentes simulations ont démontré le contraire, en restant encore un problème théorique. L'utilisation du taux de l'inertie (pourcentage de la variance) pour la "qualité de la représentation", proposé de Benzécri (1979) est très difficile d'être globalisée. Enfin, pour mesurer la variation de l'information, Kullback (1959) a proposé la théorie de Shannon-Wiener pour utiliser un indice de divergence dans les problèmes d'inférence pour la statistique multivariée (Lebart et autres, 1995). Basée sur le théorème de Bayes, on mesure la distance $J(H_1 : H_2)$ entre les hypothèses H_1 versus H_2, en tenant compte de la matrice de covariance (Γ) et des valeurs propres(λ_i) : $J(I,\Gamma) = \frac{1}{2}\left(\sum_{i=1}^{p}\lambda_i + \sum_{i=1}^{p}\frac{1}{\lambda_i}\right) - p > 0,$

où I(1 : 2) est l'information moyenne après la discrimination dans les deux espaces des deux échantillons, sous les deux hypothèses contraires. (Lebart et colab., 1995)
Si les deux inerties théoriques totales sont égales sous les deux hypothèses H_1 et H_2, alternatives,

on a : $\sum_{i=1}^{p} \lambda_i = \mathrm{p}$, donc le seul terme qui peut avoir une variation est : $\sum_{i=1}^{p} \frac{1}{\lambda i_i}$,

ce que pose en question les petites valeurs propres, tandis que l'analyse factorielle ne retient en général que les plus grandes (Lebart et collab., 1995). D'après les auteurs, on peut aussi construire un domaine critique, basé sur la théorie de Kullback (1959) :

$$\left\{ \left| \sum_{\alpha=1}^{p} \frac{1}{\lambda} - \sum_{\alpha=1}^{p} \lambda_\alpha \right| \right\} \geq C .$$

On peut observer que la divergence entre les deux hypothèses augmente dans le cas où les valeurs propres de la matrice Γ se rapprochent de zéro. Conformément à la théorie de l'information, les valeurs propres infiniment petites auront un impact plus grand par rapport à celles qui peuvent expliquer près de 80% de l'inertie totale du sous-espace des deux facteurs mis en correspondance, avec une perte considérable d'information. (Lebart et collab., 1995)

4 Indicateur informationnel agrégé

En général, dans l'analyse factorielle: $\tau_j = \lambda_j / \sum_{j=1}^{p} \lambda_j$, avec: $\sum_{j=1}^{p} \tau_j = 1$,

représente l'indicateur de la "qualité de la représentation" (Benzécri, 1979). En tenant compte de ces considérations, aussi que d'autres théories informationnelles, de Kullback (1959) et de Octav Onicescu (1963), pour une évaluation globale de l'analyse factorielle, je propose dans le présent travail un indicateur informationnel agrégé *SYNAPSe* (Popescu-Spineni, 1998), d'homogénéité-hétérogénéité, que j'ai construit, en partant des deux rapports (informationnels), construits dans ce but (voir aussi Popescu-Spineni, 2000) :

$$E(\tau) = \sum_{j=1}^{K} \frac{\lambda_j^2}{(\sum_{j=1}^{K} \lambda_j)^2} \; ; \quad (1) \qquad \text{et} \qquad E(\tau') = \sum_{j=1}^{K} \frac{\frac{1}{\lambda_j^2}}{(\sum_{j=1}^{K} \frac{1}{\lambda_j})^2} \; ; \qquad (2)$$

Ces rapports m'ont conduit à élaborer le *SYNAPSe* ou " l'*indicateur synergique*" (ou domaine critique), étant en mesure d'évaluer la signification globale de la représentation, par comparaison avec le nombre de variables retenues (1/ k, avec k < p):

(SYNAPSe): $$\boxed{\left\{ E(\tau) - E(\tau') \right\} = \frac{1}{k}} \qquad (3)$$

L'indicateur E(τ), est mis en balance, par l'indicateur *SYNAPSe*, avec E(τ'), en tenant compte qu'ils ont une variation informationnelle, stricte entre 1/K et 1, parce que la constante C = 1/ K est précisément établie, *sans avoir besoin d'une simulation*. L'*Interprétation* sans avoir perte d'information: dans le cas où la valeur de *SYNAPSe* > 1/K, il y a une signification statistique de l'analyse (H_0 acceptée) ; au contraire, pour le *SYNAPSe* ≤ 1/K, l'analyse globale n'est pas significative (H_0 rejetée). (Popescu-Spineni, 2000)
J'ai construit cet indicateur sur la base du concept *d'énergie informationnelle* (O. Onicescu, 1963), avec la définition (4) suivante :
Définition : L'information globale du système S avec les états $s_1, s_2, \ldots, s_n$ ayant les pondérations $p_1, p_2, \ldots, p_n$ peut-être exprimée par *l'énergie informationnelle* :

$$E = \sum_{i=1}^{n} p_i^2 = 1 \ , \text{ avec : } \sum_{i=1}^{n} p_i = 1 \quad \text{et} \quad \frac{1}{n} \le E \le 1 \ . \tag{4}$$

On peut donner des estimations (Krippendorf, 1971) similaires pour l'indice d'hétérogénéité de C. Gini (1958), en utilisant aussi les Théorèmes de Gh. Mihoc (1980), avec des valeurs très approchés de l'énergie informationnelle (Onicescu et Stefanescu, 1979):

$$\hat{E}\ (z) = \frac{n}{n-1} \cdot E\ (z) - \frac{1}{n-1}\ , \tag{5}$$

où E(z) est l'énergie informationnelle de sélection et $\hat{E}$ est l'estimateur.
Pour n>10, on peut négliger le terme 1/ (n-1), en ayant : $\hat{E}\ (z) = \frac{n}{n-1} \cdot E\ (z)$ (6)
On peut donner pour (4) aussi des pondérations et on peut continuer avec la partie théorique.

5 Applications

5.1 Application sur un tableau de contingence "4x10"

Pour le commencement, on peut présenter l'exemple d'utilisation comparative des deux indicateurs informationnels, $E(\tau_j)$ et $E(\tau'_j)$, aussi que de *SYNAPSe*, sur les valeurs propres d'une application de A. Rizzi ("*Analisi dei Dati*", Roma, 1989, p.188), pour un tableau "4x10", (TAB. 1) :

	λ_j	τ_j	$1/\lambda_j$	$\tau'j$
1	0,0443	0,532	22,573	0,004
2	0,0202	0,242	49,506	0,008
3	0,0088	0,106	113,636	0,018
4	0,0054	0,065	185,785	0,029
5	0,0023	0,028	434,783	0,069
6	0,0021	0,025	476,191	0,076
7	0,0002	0,003	5000,000	0,796
total	0,0833	1,000	6282,474	1,000

TAB. 1 – Les valeurs propres et les inerties de l'analyse d'un tableau de contingence "4x10".

En calculant: E(τ) et E(τ') et aussi (3) l'indicateur *SYNAPSe*:

$$\boxed{\left\{\left|E(\tau)-E(\tau')\right|\right\}=\frac{1}{k}},$$

on obtient un résultat global significatif:

$E(\tau_j) = 0{,}357$; $E(\tau'_j) = 0{,}645$; où: k = 7;

$|E(\tau_j) - E(\tau'_j)| = 0{,}645 - 0{,}357 = 0{,}288 > 0{,}143 = \frac{1}{7}$.

En conclusion, il faut rejeter H_0, car le tableau n'est pas bien équilibré.

5.2 Application sur deux tableaux de contingence "41x15"- cas (a) et cas (b)

Dans un ancien projet de recherche, j'ai fait un étude sur la situation de la répartition par districts (41 districts), pour 15 spécialités médicales, des médecins de la Roumanie de l'an 1985 ((a)-données en chiffres absolues (CA), (b)-données rapportées à 10000 habitants ($^0/_{0000}$)), en utilisant l'analyse des correspondances, aussi que l'analyse cluster, en parallèle, sur les deux tableaux de contingence de "41x15" (TAB. 2). A l'aide de l'analyse des correspondances, faite sur la répartition des médecins par districts, pour les 15 spécialités, dans le cas (b)- données rapportées à 10000 habitants, j'ai obtenu une représentation duale sur les deux axes factoriels F1 (avec λ_1= 0,01449356) et F2 (avec λ_2 = 0,00876388), en caractérisant la variabilité totale avec 31,8% et respectivement 19,2%, ayant une qualité de la représentation de τ_j = 51,0%, où K = 14.

	(a) λ_j	τ_j (%)	**(b)** λ_j	τ_j (%)
λ_1	0,0154	42,4	0,0145	31,9
λ_2	0,0046	12,8	0,0088	19,3
λ_3	0,0042	11,6	0,0067	14,7
λ_4	0,0027	7,5	0,0037	8,2
λ_5	0,0023	6,4	0,0029	6,4
λ_6	0,0017	4,6	0,0019	4,2
λ_7	0,0012	3,4	0,0017	3,7
λ_8	0,0011	3,1	0,0014	3,2
λ_9	0,0009	2,4	0,0011	2,4
λ_{10}	0,0007	1,9	0,0009	2,0
λ_{11}	0,0005	1,4	0,0007	1,5
λ_{12}	0,0004	1,1	0,0005	1,1
λ_{13}	0,0003	0,9	0,0004	0,8
λ_{14}	0,0002	0,5	0,0003	0,6
λ_{15}	0,000	0,0	0,000	0,0
total	φ = 0,03619	100.	φ = 0,04551	100.

TAB. 2 – Les valeurs propres et les inerties de l'analyse des deux tableaux "41x15"- cas (a) et cas (b).

Pour le cas (a), il faut calculer les indicateurs (1) $E(\tau_j)$ et (2) $E(\tau'_j)$,
nous avons: $E(\tau_j) = 0{,}223$; $E(\tau'_j) = 0{,}146$;
Avec l'indicateur *SYNAPSe* : $|E(\tau_j) - E(\tau'_j)| = 0{,}223 - 0{,}146 = 0{,}077 \geq 0{,}071 = 1/14$,
on obtient un résultat significatif (à la limite- cas des médecins spécialistes en valeurs absolues), en montrant une très forte corrélation avec la première axe factoriel F1, donc avec les anciens cinq centres universitaires du notre pays. (Popescu-Spineni, 1998; 2000).

Pour le cas (b), il faut calculer les indicateurs (1) $E(\tau_j)$ et (2) $E(\tau'_j)$,
nous avons: $E(\tau_j) = 0{,}179$ et $E(\tau'_j) = 0{,}136$; où: $k = 14$;

	(a) λ_j	τ_j	(a) $1/\lambda_j$	τ'_j
λ_1	0,01536	0,424	65,104	0,003
λ_2	0,00462	0,128	216,450	0,011
λ_3	0,00420	0,116	238,095	0,012
λ_4	0,00272	0,075	367,647	0,019
λ_5	0,00231	0,064	432,900	0,023
λ_6	0,00165	0,046	606,060	0,032
λ_7	0,00125	0,034	800,000	0,042
λ_8	0,00113	0,031	884,000	0,046
λ_9	0,00085	0,024	1176,000	0,061
λ_{10}	0,00068	0,019	1420,588	0,076
λ_{11}	0,00052	0,014	1923,077	0,100
λ_{12}	0,00039	0,011	2564,102	0,133
λ_{13}	0,00031	0,009	3225,807	0,168
λ_{14}	0,00019	0,005	5263,158	0,274
total	0,03619	1,000	19232,98	1,000

TAB. 3 – Les valeurs propres et leurs inverses de l'analyse d'un tableau de contingence "41x15"- cas (a).

Avec l'indicateur *SYNAPSe* : $|E(\tau_j) - E(\tau'_j)| = 0{,}179 - 0{,}136 = 0{,}043 < 0{,}071 = 1/14$,
on obtient un résultat non-significatif. Il faut accepter H_0, car l'analyse n'a pas montré des directions privilégiées (cas des médecins spécialistes par habitants ($^0/_{0000}$)). (Popescu-Spineni, 1998; 2000). Les valeurs propres (λ_j) et les $1/\lambda_j$ du cas (a) en valeurs absolues sont dans TAB. 3.

La représentation duale a montré une agglomération des «objets » (districts) et des « variables » (spécialités) autour l'origine des axes factoriels, plus accentuée dans l'analyse (b) que dans l'analyse (a), due au rapport des spécialistes par habitants ($^0/_{0000}$). Autour du centre des axes, l'inertie du nuage est plus faible sans avoir des directions privilégiées. Dans l'analyse (a) une forte corrélation des districts avec les centres universitaires a été observée sur le premier axe F1, qui ne se maintient que partiellement dans l'analyse (b), où la distribution des districts et des

spécialités est plus équilibrée autour du centre des axes. Cette situation, plus équilibrée dans le cas (b)- pour $^{0}/_{0000}$, montrée par comparaison avec le cas (a), se vérifie avec les indicateurs informationnels décrits plus haut, surtout avec l'indicateur synergique *SYNAPSe*, appliqués sur les valeurs propres et leurs inverses, des deux tableaux de contingence "41x15" du modèle.

6 Conclusions

Par l'intermède de la statistique informationnelle, surtout à l'aide du concept de l'énergie informationnelle (Onicescu, 1968), cette étude propose deux indicateurs d'homogénéité ou hétérogénéité, $E(\tau_j)$ et $E(\tau'_j)$, construits pour les valeurs propres λ_j et $1/\lambda_j$ (j = 1, …, K) et aussi, l'indicateur synergique, *SYNAPSe* ou domaine critique, par rapport à la valeur C = 1/K, c'est-à-dire le domaine des valeurs de sélection pour lesquelles on rejette l'hypothèse nulle H_0. Il a résulté ainsi une méthode pratique pour réduire la dimension (p < K) de l'espace factoriels des axes, qui est le but des diverses modèles d'analyse factorielle. (Popescu-Spineni, 1998)

Dans ce but, l'indicateur $E(\tau)$, a été mis en balance avec $E(\tau')$, par l'indicateur agrégé *SYNAPSe*, en tenant compte qu'ils ont une variation informationnelle, stricte entre 1/K et 1, car la constante C = 1/ K est précisément établie. On peut ainsi optimiser l'interprétation globale de la qualité de la représentation, sans avoir perte d'information: dans le cas où la valeur de *SYNAPSe* > 1/K, il y a une signification statistique de l'analyse globale (H_0 acceptée); au contraire, pour *SYNAPSe* ≤ 1/K, l'analyse n'est pas significative (H_0 rejetée). Sans une utilisation appropriée de l'énergie informationnelle, le domaine critique pourrait être obtenu seulement par simulation.

Références

Benzécri, J.P. (1980) *L'Analyse des données.* Paris : Dunod.

Lebart, L., A. Morineau, M. Piron (1995) *Statistique exploratoire multidimenionnelle.* Paris : Dunod.

Onicescu, O., V. Stefanescu (1979) *Elemente de statistica informationala cu aplicatii.* Bucuresti : Editura Tehnica.

Popescu-Spineni, S. (2000) *Tehnici de* a*naliza datelor multidimensionale – MDA.* Bucuresti : Editura Universitara UMF "Carol Davila".

Popescu-Spineni, S. (1998) *Hierarchy Techniques of Multidimensional Data Analysis (MDA) in Social Medicine Research*. Roma: Studies in Classification, Data Analysis and Knowledge Organization. Berlin: Springer Verlag (A. Rizzi, M. Vichi, H.H. Bock Editors), 641-646.

Rizzi, A. ed. (1995) *Some Relation between Matrices and Structures of Multi-dimensional Data Analysis.* Pisa: Giardini Editori e Stampatori.

Saporta, G., V. Stefanescu (1996) *Analiza datelor & informatica.* Bucuresti: Editura Economica.

[1] SYNAPSe (ou SYNPS) c'est le même indicateur.

Summary

The main goal of factorial analysis is to reduce the dimensions of a spatial (dual) representation of observations, by reducing the number of factorial axes. It is very difficult to establish a good evaluation criterion for the representation quality in the reduced space, after having reduced the factorial axes. There are many empirical techniques for reducing the number of factorial axes (Kaiser, Cattel), but this study will present a criterion for finding and testing an informational aggregate indicator (Popescu-Spineni, 1998), which has been built by the intermediate of an informational statistical concept, without loss of information. Examples are proposed for applying and testing the described informational aggregate indicator, based on theoretical references.