Indicateur informationel aggrégé pour évaluer la qualité de l'analyse des données sans avoir manque d'information

Sabina Popescu – Spineni
Institut de Santé Publique & Université de Médicine
et Pharmacie "Carol Davila" - Bukarest, Roumanie.
annalyse factorielle, statistique informationnelle, santé publique
spopescu@ispb.ro

1. Introduction

Dans les sciences médico-sociales, aussi que dans le management sanitaire, l'analyse multivariée des données a été imposée par des besoins opérationels. Mais l'analyse des données utilise beaucoup de méthodes d'optimisation qui proposent des algorithmes très rigoureuses pour établir la partition d'un set *n* d'objets caractérisés par *k* variables, qui charactérise 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 linéaire des données (composantes principales, cannonique, factorielle, analyse des correspondences), simples ou multiples. On se propose une représentation synthétique avec le minimum de manque d'information.

Le but de l'analyse factorielle comme téchnique d'analyse multivariée est de déterminer si la corrélation d'un grand nombre de variables obsérvées peut *ê*tre expliquée par un petit nombre de facteurs fondamentales et combien de tels facteurs on a besoin. Des méthodes classiques d'optimisation sont adaptées pour suivre le type des donées. [1]

Les solutions des méthodes méntionnées sont appropriées pour déterminer un nombre minimal de facteurs capables d'expliciter toute la variabilité, par une symmétrie parfaite entre les profiles-colonnes et les profiles-lignes analysées, en considérant les constraintes et en envisgeant les priorités. Le concept de *distance* joue un rolle essentiel dans la plusparte des modèlles d'analyse multivariée, en considerant les différentes diréctions de dispersion de la variabilitée, mais la théorie des *valeurs propres* prends la plus grande importance 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 d'axes factorielles. Mais il est trés difficile d'établir un critère adéquate pour mesurer la qualité de la représentation dans l'éspace factoriel réduit, après ayant réduire le nombre des axes factorielles. [5]

On a beaucoup de téchniques empiriques pour réduire le nombre des axes factorielles (Kaiser, Cattel), mais dans cet étude on présente un bon critère pour trouver et tester un indicateur informationel agreggé (Popescu-Spineni, 1998), celui-çi calculé par l'aide d' un concept statistique informationel, sans avoir manque d'information. Sont bien proposés beaucoup d'éxemples, pour appliquer et tester cet indicateur informationel aggrégé, avec des réfférences théoriques appropriées.

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

La réduction de la dimension de l'éspace (cartésien) des obsérvations dans l'analyse factorielle ou des composantes principales revient à la réduction du nombre des axes factorielles, jusqu'à présent, sans une solution rigoureuse.

Le but de l'analyse étant d'obtenir une représentation des obsérvations dans l'éspace de dimension réduite (soit < p), on doit apprécier la p*ê*rte d'information après avoir calculé le nombre des facteurs retenus. On obsèrve dans ce cas que le nuage des points n'est pas encore centré dans le centre de gravité (G). Après avoir calculé les valeures propres (λ_i, avec $1 \leq i \leq p$) du système des k écuations, on considère le critère du pourcentage de l'inertie totale (τ) comme mesure pour apprécier 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 éxemple, avec $(\lambda_1 + \lambda_2)/(\varphi^2) = 0{,}9$, donc 90%, on rézulte que le nuage des points est applatisé sur un souséspace avec deux dimensions (Saporta, 1996). On peut obsérver que, dans l'appréciation du pourcentage de l'inertie, on doit tenir compte du nombre des variables initiales: un pourcentage de 10% n'a pas la même signification pour un tableau avec 20 variables comme pour un autre de 100 variables. [1]

Ont été élaborés des critères théoriques (Saporta, 1996) qui proposent déterminer si les valeures propres sont significatives différentes, en commençant d'un certain rang, mais si non, on retient les premières valeures propres (λ_i). La réduction de dimension n'est pas possible seulement s'il éxiste une rédondance entre les variables $x_1, \ldots, x_p$, ou non, s'il sont indépendantes (Saporta). Par la méthode des approximations succésives, on éstime les valeures propres et les vecteurs propres associés à l'équation caractéristique $|R-\lambda I| = 0$, qui éxplique la plusparte de la variation totale. En suposant que le processus c'est interrompu apres avoir éstimer p ($p < K$) valeurs propres, on doit vérifier si les plus petites des valeurs propres sont égales, pour maximiser la variance des composantes principales de l'analyse. [1]

En réalité, on s'applique des critères empiriques (Kaiser, 1965), étant le cas par éxemple, de l'analyse des composantes principales pour les valeurs propres supérieures à 1.

En général, pour résoudre l'équation statistique matricéale $Y = AX + \varepsilon$ dans l'analyse factorielle, on doit proposer un critère d'optimisation et la méthode d'éstimation des paramètres. Pour l'éstimation optimale des paramètres, il existent deux conditions principales: la réproduction des corrélations obsérvées et l'éxplication de la variabilité (Torrens-Ibern, 1972). Les méthodes statistiques pour l'éstimation des paramètres dépendent de la technique de résoudre le système d'équation linéaires: (1) la méthode des moindres quarées; (2) le principe de la vérosimilitée maximalle (maximum likelihood); - les deux étant équivalentes dans le cas normal (Gauss-Laplace).

En général, dans l'analyse factorielle, le problème de tester le nombre des facteurs (ou des axes) qu'on doit retenir se pose de deux diréctions: soit on supose l'éxistence d'un nombre p de facteurs communs en vérifiant l'hypothèse nulle H_0, soit on détermine le nombre intégral des facteurs, en calculant leurs nombre en partant du nombre des obsérvations. [2]

Les plus importantes critères théoriques proviennent de Barttlet (1951) et Lawley (1940), mais Jöreskog (1963) propose une procédure spécifique pour réduir le nombre des facteurs dans le cas des obsérvations non-normales ("sans hypothèses préalables")[1]

3. L'aproximation de la distribution des valeures propres

La réduction du nombre des 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'indépendence des lignes et des colonnes, cette distribution s'approche d'une lois de distribution connue d'une matrice Wishart (Lebart, Morineau, Piron, 1995). On se souligne que la lois de distribution des valeurs propres résultées de l'analyse a été erroné appréciée, aussi que l'inertie totale du nuage des points, comme de type χ^2, pendant que différentes simulations ont démontré au contraire, en restant un problème théorique encore. 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 divérgence 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 (altérnatives), 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 éspaces des deux échantillons, sous les deux hypothèses contraires (Lebart et autres, 1995).
Si les deux inerties théoriques totales sont égales sous les deux hypothèses H_1 et H_2, on a:

$$\sum_{i=1}^{p}\lambda_i = p$$

Donc, le seul tèrme qui peut avoir une variation est: $\sum_{i=1}^{p}\frac{1}{\lambda i_i}$,

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

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

On peut obsérver que la divérgence entre les deux hypothèses augmente dans le cas où les valeurs propres de la matrice Γ se rapprochent de zéro. Conformement au théorie de l'information, les valeurs propres infiniment petites aurront un impact plus grand par rapport aux celles qui peuvent éxpliquer près de 80% de l'inertie totale du souséspace des deux facteurs mis en correspondence, avec une pêrte considérable d'information (Lebart et autres, 1995). [2]

4. Indicateur informationel aggrégé :

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 Onicescu, 1963), pour une évaluation globale de l'analyse factorielle, je propose dans le présent travail un indicateur informationel aggrégé (IS-Popescu-Spineni, 1998), d'omogénité-hétérogénité, que moi même j'ai construit, en partant des deux rapports (informationels), construites dans ce but: [3] [4]

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

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

$$(ISPS): \qquad \{|E(\tau) - E(\tau')|\} \geq \frac{1}{K}, \qquad (3)$$

L'indicateur $E(\tau)$, est mis en balance (dans *ISPS*) avec l'indicateur, $E(\tau')$, en tenant compte qu'ils ont une variation informationnelle, stricte entre 1/K et 1, car la constante C = 1/ K est précisement établie, sans avoir besoin d'une simulation. *Interprétation,* sans avoir manque d'information: dans le cas où la

valeur de ISPS > 1/K, on a une signification statistique de l'analyse (H_0 acceptée), au contraire, pour ISPS ≤ 1/K, l'analyse globale n'est pas significative (H_0 rejetée). [3] [4]

5. Applications

5.1. Pour le commencement, je donnerai un exemple d'utilisation comparative des deux indicateurs informationels $E(\tau_j)$ et $E(\tau'_j)$, aussi que de *ISPS* (*l'indicateur synérgique*) sur les valeurs propres sorties d'une application éffectuée de A. Rizzi ("*Analisi dei Dati*", Roma, 1989, p.188), sur un tableau "4x10":

Exemple No.1:

Ex.1	λ_j	τ_j	cum.	cum.	$1/\lambda_j$	$\tau'j$
1	0,0443	0,532	|	1,000	22,573	0,004
2	0,0202	0,242	0,532	0,996	49,506	0,008
3	0,0088	0,106	0,772	0,998	113,636	0,018
4	0,0054	0,065	0,887	0,970	185,785	0,029
5	0,0023	0,028	0,942	0,941	434,783	0,069
6	0,0021	0,025	0,970	0,872	476,191	0,076
7	0,0002	0,003	0,997	0,796	5000,000	0,796
	0,0833	1,000	1,000	–	6282,474	1,000

En calculant:

(1) $E(\tau_j) = \sum_{j=1}^{K} \frac{\lambda_j^2}{(\sum_{j=1} \lambda_j)^2}$; et (2) $E(\tau'_j) = \sum_{j=1}^{K} \frac{1/\lambda_j^2}{(\sum_{j=1} 1/\lambda_j)^2}$;

En calculant aussi l'indicateur *ISPS*: $\{|E(\tau)-E(\tau')|\} \geq \frac{1}{K}$, (3)

on obtient un résultat global significatif:

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

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

En conclusion, il faut rejeter H_0, car le tableau n'est pas équilibré. [3] [4]

5.2. Dans un autre 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 ciffres absolues, (b)-donnees rapportées pour 10000 habitants), en utilisant l'analyse des correspondances, aussi que l'analyse cluster, en parallel, sur un tableau de contingence de "41x15". 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 par 10000 habitants, j'ai obtenu une représentation duale sur les deux axes factorielles 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 = 15 (j = 1, …, K). [3] [4]

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

partiallement dans l'analyse (b), où la distribution des districts et des spécialités est plus équilibrée autour du centre des axes. Cetté situation plus équilibrée montrée par l'analyse duale dans le cas (b) se vérifie avec les indicateur informationels décrits plus haut, surtout avec *l'indicateur synérgique ISPS*, appliqués sur les valeurs propres (cas (b)- pour $^0/_{0000}$), sorties de l'analyse d'un tabeau "41x15" :

Exemple No. 2 :

	A1 **(a)** λ_j	inerție(%)	A2 **(b)** λ_j	inerție (%)
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}	0,0000	0,5	0,0000	0,0
	φ = 0,03619	100.	φ = 0,04551	100.

(b) En calculant les indicateurs (1) $E(\tau_j)$ et (2) $E(\tau'_j)$,
on obtient: $E(\tau_j)$ = 0,179 et $E(\tau'_j)$ = 0,136; où: k = 14;
En calculant aussi l'indicateur *ISPS*, on obtient un résultat non-significatif:
$\{|E(\tau)-E(\tau')|\}$ = 0,179 - 0,136 = 0,043 < 0,071 = 1/14.

Il faut accepter H_0, car l'analyse n'a pas montré des diréctions privilégiées, pour le cas des médecins spécialistes rapportés par habitants ($^0/_{0000}$). [3] [4]

Références:

1. Benzécri, J.P. – *L'Analyse des donnees*, Dunod, Paris, 1980;
2. Lebart, L., Morineau, A., Piron, M.– *Statistique exploratoire multidimensionnelle*, Dunod, Paris, 1995;
3. Popescu-Spineni Sabina – *Statistical Methods for Multidimensional Data Analysis and Classification*; Editura Universitara "Carol Davila", 2000, (145 pages);
4. Popescu-Spineni Sabina - *Hierarchy Techniques of Multidimensional Data Analysis (MDA) in Social Medicine Research*; Studies in Classification, Data Analysis and Knowledge Organization, Springer Verlag (A. Rizzi, M. Vichi, H.H. Bock Editors), 641-646, 1998;
5. Rizzi, A., ed. – *Some Relation between Matrices and Structures of Multidimensional Data Analysis*; Giardini Editori e Stampatori, Pisa, 1995.