Statistique et analyse des données

Pablo Gonzalez Vicente

Recherche de la dimension de l'espace latent en ACP

Statistique et analyse des données, tome 11, n° 3 (1986), p. 19-29
<http://www.numdam.org/item?id=SAD_1986__11_3_19_0>

Statistique et Analyse des Données
1986 - vol.11 n°3 pp.19-29

RECHERCHE DE LA DIMENSION DE L'ESPACE LATENT EN ACP

GONZALEZ VICENTE Pablo

Electricité de France
Direction des Etudes et Recherches

Résumé : *On étudie une condition nécessaire et une condition suffisante de positivité pour des matrices symétriques et on donne une interprétation géométrique de ces résultats. On propose ensuite une interprétation de l'ACP et un critère simple pour la recherche de la dimension de l'"espace latent".*

Abstract : *In this paper we give one sufficient condition and one necessary for symetric matrices to be positive. We also give a geometrical interpretation to these results. Finally we suggest a new way of considering Principal Compenent Analysis and a simple criterion for determining the latent space dimension.*

Indices de classification STMA : *06-070, 06-080.*

1 - GENERALITES

Soit S(n) l'espace vectoriel réel des matrices symétriques n x n à coefficients réels. Nous fournissons cet espace du produit interne :

$$\langle A,B \rangle = \text{Trace } (AB') \qquad A,B \in S(n)$$

et de la norme associée :

$$|A| = (\text{Trace } (AA'))^{\frac{1}{2}}$$

Manuscrit reçu le 22 juin 1986
Révisé le 1 juin 1987

Si τ dénote le sous-ensemble de S(n) des matrices symétriques et définies (semi)-positives, alors :

* τ est un cône convexe fermé avec sommet à l'origine,
* $\overset{\circ}{\tau}$ est l'ensemble des matrices strictement définies positives,
* Frontière (τ) est l'ensemble des matrices semi-définies positives de rang non complet.

Rappelons finalement que pour toute matrice $A \in S(n)$ existe P orthogonale tel que : $P' AP = D$ D matrice diagonale $PP' = P'P = I$

et, en plus : $|A| = |D|$, trace (A) = trace (D).

2 - CONDITIONS DE SEMI-POSITIVITE

Théorème 1 :

Soit $A \in S(n)$

Si $\frac{\text{Trace } (A)}{|A|} \geq \sqrt{n-1}$ alors A est semi-définie positive

L'inégalité stricte implique la stricte positivité de A

Démonstration :

Soit $A \in S(n)$ tel que : $\frac{\text{Trace } (A)}{|A|} \geq \sqrt{n-1}$

Si A n'est pas semi-définie positive alors A à une valeur propre négative ; soit celle-ci λ_1.

Soit $D^* = \begin{bmatrix} 0_1 & & \\ & \ddots & \\ & & 1 \end{bmatrix}$ et $A = PDP'$, $D = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}$ avec $\lambda_1 < 0$.

Alors : $\quad <I-D^*, D> = \lambda_1 < 0$

$\Rightarrow \quad <I, D> < <D^*, D> \leq |D^*|\ |D|$

Donc : $\frac{\text{Trace (A)}}{|A|} = \frac{\text{Trace (D)}}{|D|} < |D^*| = \sqrt{n}-1$

ce qui contredit l'hypothèse du départ.

Théorème 2 :

Soit $A \in S(n)$
Si A est semi-définie positive alors $\frac{\text{Trace (A)}}{|A|} \geq 1$

Démonstration :

Soit A semi-définie positive ; alors :

$A = PDP'$, $D = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}$ avec $\lambda_i \geq 0 \quad \forall_i = 1,\ldots,n$

Soit $\{D_i\}_{1 \leq i \leq n}$ la base canonique des matrices diagnonales de dim.n

Alors $|A| = |D| = |\sum_{i=1}^{n} \lambda_i D_i| \leq \sum_{i=1}^{n} \lambda_i |D_i| = \sum_{i=1}^{n} \lambda_i = \text{Trace (D)} = \text{Trace (A)}.$

Donc : $\frac{\text{Trace (A)}}{|A|} \geq 1.$

Corolaire :

Soit $A \in S(n)$; alors :

* Si n=2, A est semi-définie positive $\Leftrightarrow \frac{\text{Trace (A)}}{|A|} \geq 1$

* Si rang (A) = K alors A semi-définie positive $\Rightarrow \frac{\text{Trace (A)}}{|A|} \geq \sqrt{K}$

3 - INTERPRETATION GEOMETRIQUE

Remarquons que :

$$\frac{\text{Trace }(A)}{|A|} \geqslant \sqrt{n-1} \Leftrightarrow \frac{\langle A,I\rangle}{\sqrt{n}\,|A|} \geqslant \frac{\sqrt{n-1}}{\sqrt{n}} \Leftrightarrow \frac{\langle A,I\rangle}{|A|\,|I|} \geqslant \frac{\sqrt{n-1}}{\sqrt{n-1}\,\sqrt{n}}$$

Notons $E_1 = I - D_1 = \begin{bmatrix} 0_1 & & \\ & \ddots & \\ & & 1 \end{bmatrix}$ alors :

$$\frac{\text{Trace }(A)}{|A|} \geqslant \sqrt{n-1} \Leftrightarrow \frac{\langle A,I\rangle}{|A|\,|I|} \geqslant \frac{\langle E_1,I\rangle}{|E_1|\,|I|} \Leftrightarrow \text{angle } (A,I) \leqslant \text{angle } (E_1,I).$$

Le théorème 1 se réecrit donc :

$A \in S(n)$ et angle $(A,I) \leqslant$ angle $(E_1,I) \Rightarrow$ A semi-définie positive.

D'autre part :

$$\frac{\text{Trace }(A)}{|A|} \geqslant 1 \Leftrightarrow \frac{\langle A,I\rangle}{|A|\,|I|} \geqslant \frac{1}{|I|} = \frac{|D_1|}{|D_1|\,|I|} = \frac{\langle D_1,I\rangle}{|D_1|\,|I|}$$

Le théorème 2 se réecrit donc :

$A \in S(n)$ et A semi-définie positive $\Rightarrow$ angle $(A,I) \leqslant$ angle (D_1,I).

Nous pouvons conclure que le cône $\mathcal{C}$ des matrices semi-définie positives est contenu dans un cône de révolution maximal, dont la génératrice est la matrice D_1 et l'axe de rotation la matrice identité I.

Evidemment toutes les matrices dans ce cône-là ne sont pas semi-définies positives, mais il existe un cône de révolution minimal, dont la génératrice est la matrice $E_1 = 1 - D_1$ et l'axe de rotation est l'identité I, tel que toutes les matrices contenues dans lui sont semi-définies positives.

Dans la figure 1 nous donnons cette interprétation pour les matrices diagonales de 3 x 3 :

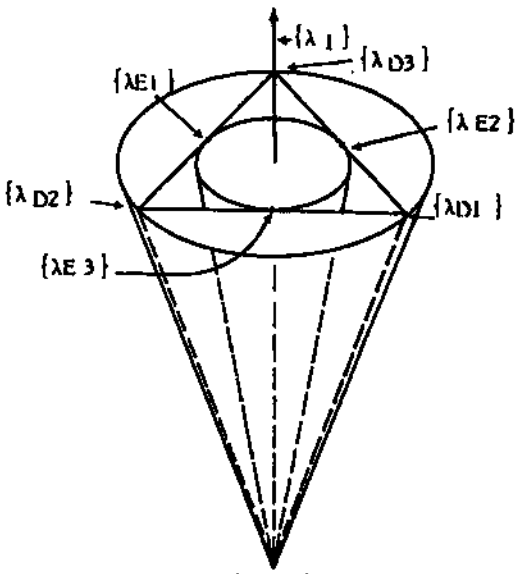


Figure 1

Une interprétation géométrique très intéressante peut être donnée dans S(2) grâce à l'identification :

$$\begin{bmatrix} X & Z \\ Z & Y \end{bmatrix} \longleftrightarrow \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \in \mathbb{R}^3$$

Dans ce cas particulier $\mathcal{C}$ est exactement le cône de révolution :

$$\{X \quad S(2) \text{ tel que angle } (X,I) \leq \text{angle } (D_1,I) = 45°\}$$

La classe de similarité $\mathcal{C}(D)$ associée à une matrice diagonale D $\mathcal{C}(D) = \{ X \in S(2)$ tel que $X = PDP'$ avec P orthogonale $\}$ est une circonférence qui a pour centre une matrice proportionnelle à l'identité.

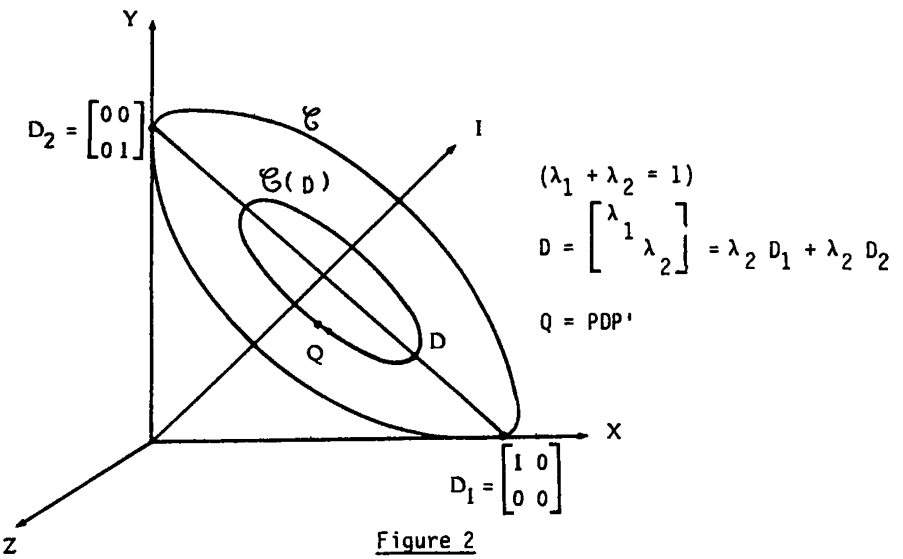


Figure 2

Toute matrice semi-définie positive peut être interprétée alors comme une "rotation" autour de l'axe $\{ \lambda I \ ; \ \lambda \geq 0 \}$ d'une matrice diagonale D (à composantes non négatives), dont les paramètres de cette rotation sont donnés par la matrice orthogonale P.

4 - INTERPRETATION DE L'ACP

Soit X un tableau de données à q variables et n individus et X'X la matrice de covariances empiriques (nous supposons les variables centrées) ; alors $X'X \in \mathcal{C} \subset S(q)$.

Nous pouvons maitenant associer à X'X une base de S(q), où sa représentation est particulièrement portant :

$$X'X = PDP' = P \left(\sum_{i=1}^{q} \lambda_i D_i \right) P' = \sum_{i=1}^{q} \lambda_i P D_i P' = \sum_{i=1}^{q} \lambda_i Q_i \text{ avec } Q_i = PD_i P'$$

La base $\{Q_i\}_{1 \leq i \leq q}$ est une base orthonormée et chaque Q_i peut être écrit en fonction des vecteurs propres P_i de X'X : $Q_i = P_i P'_i$.

Dans l'espace S(2) l'interprétation géométrique est évidente (figure 3) :

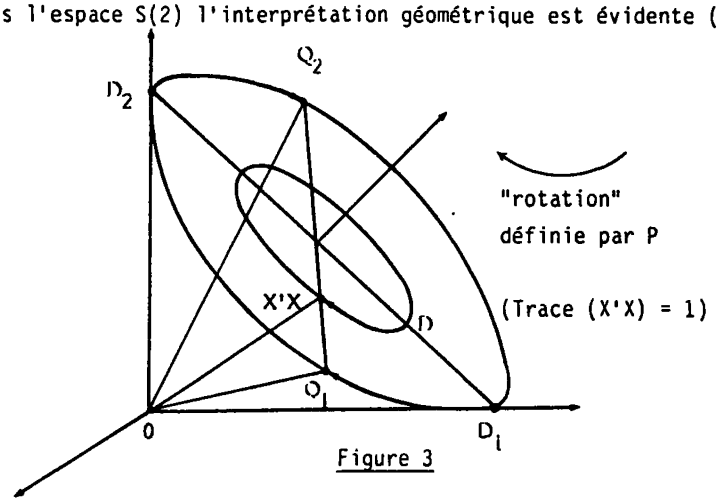


Figure 3

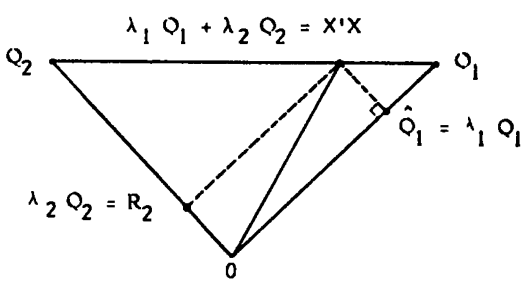
Retenir les n premiers axes dans l'ACP revient à projeter la matrice X'X sur le sous espace vectoriel engendré par les matrices Q_i associés aux n plus grandes valeurs propres λ_i. Etant donné que $\{Q_i\}_{1 \leq i \leq q}$ est une base orthonormée, cette projection est :

$$\hat{Q}_n = \sum_{i=1}^{n} \lambda_i Q_i \qquad \text{rang } (\hat{Q}_n) = n$$

et le "résidus" R_n :

$$R_n = Q - \hat{Q}_n = \sum_{i=n+1}^{q} \lambda_i Q_i \qquad \text{avec } ||Q - \hat{Q}_n||^2 = \sum_{i=n+1}^{q} \lambda_i^2 \text{ et } \hat{Q}_n \perp Q - \hat{Q}_n$$

Reprenant la fig. 3 nous avons :



Le problème qui se pose tout naturellement est le choix optimal de n dans la décomposition $X'X = \hat{Q}_n + R_n$.

Si l'on interprète $\hat{Q}_n$ comme la matrice de variance expliquée par les premiers n facteur, il est naturel de supposer que la matrice de variance de résidus soit de la forme $R_n = \sigma^2 I$. Evidemment cette condition sera difficilement réalisée dans la pratique. Mais lorsque l'on utilise l'ACP comme approximation du modèle factoriel, on peut chercher n, de façon que cette condition soit le mieux satisfaite.

La proximité entre R_n et $\sigma^2 I$ est mesurée par le cosinus de l'angle θ entre ces deux matrices :

$$\cos\theta\ (R_n, \sigma^2 I) = \frac{< R_n, I>}{|R_n|\ |I|} = \frac{\text{Trace}\ (R_n)}{q\ |R_n|}$$

Le critère sera donc :

$$\underset{n}{\text{Max}} \quad \frac{\sum_{i=n+1}^{q} \lambda_i}{(\sum_{i=n+1}^{q} \lambda_i^2)^{\frac{1}{2}}} \quad \Leftrightarrow \quad \underset{n}{\text{Min}} \quad \theta\ (R_n, I)$$

5 - APPLICATION

Nous reprenons ici les exemples donnés par ANASTASSAKOS et D'AUBIGNY (1) ; nous arrivons au mêmes résultats avec un critère plus simple que le test de sphérécité empirique (VELICER) pour eux exemplifié.

Cas 1 :

On considère le modèle :

$$Y_j = \sum_{i=1}^{3} a_{ij} X_i = u_j \qquad j = 1, \ldots, 6$$

Les variables X_i ($1 \leqslant i \leqslant 3$) ont été générées de 3 échantillons gaussiens indépendants ($N = 100$, $m = 0$, $\sigma = 1$) et les u_j ont été générés de 6 échantillons indépendants (entre eux et des échantillons précédents) gaussiens ($m = 0$, $\sigma = 0.01$). Les coefficients sont :

$$y_1 = .5X_1 + .3X_2 + .2X_3$$
$$y_2 = .7X_1 + .2X_2 + .1X_3$$
$$y_3 = .9X_1 + .8X_2 + .02X_3$$
$$y_4 = .3X_1 + .1X_2 + .6X_3$$
$$y_5 = .01X_1 + .8X_2 + .19X_3$$
$$y_6 = .2X_1 + .35X_2 + .45X_3$$

Matrice de correlations =

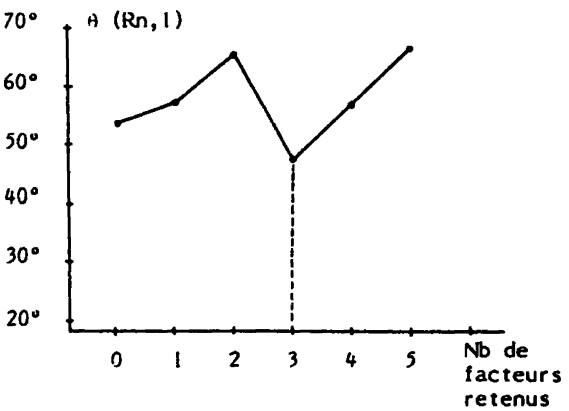
$$\begin{bmatrix} 1 & & & & & \\ .928 & 1 & & & & \\ .821 & .964 & 1 & & & \\ .677 & .532 & .416 & 1 & & \\ .448 & .150 & -.066 & .215 & 1 & \\ .752 & .502 & .309 & .872 & .645 & 1 \end{bmatrix}$$

L'ACP fournit les valeurs propres :

$\lambda_1 = 3.874$ $\lambda_2 = 1.425$ $\lambda_3 = 0.675$
$\lambda_4 = 0.012$ $\lambda_5 = 0.009$ $\lambda_6 = 0.006$

L'angle entre $R_n = X'X - \hat{Q}_n$ et I :

Nb de facteurs retenus :	0	1	2	3	4	5
$\theta(R_n, I)$:	54°	57°	65°	47°	56°	66°



Cas 2 :

Nous étudions maintenant une situation d'équicorrélation :

Matrice de correlations =

$$\begin{bmatrix} 1 & P & \cdots & P \\ P & 1 & \cdots & P \\ \vdots & \vdots & & \vdots \\ P & P & \cdots & 1 \end{bmatrix} \qquad 0 < P \leqslant 1$$

L'ACP de cette matrice fournit les valeurs propres :

$$\lambda_1 = 1 + (q-1)P \qquad \lambda_2 = \ldots = \lambda_q = 1-P$$

Calcul de $\theta\ (R_n, I)$:

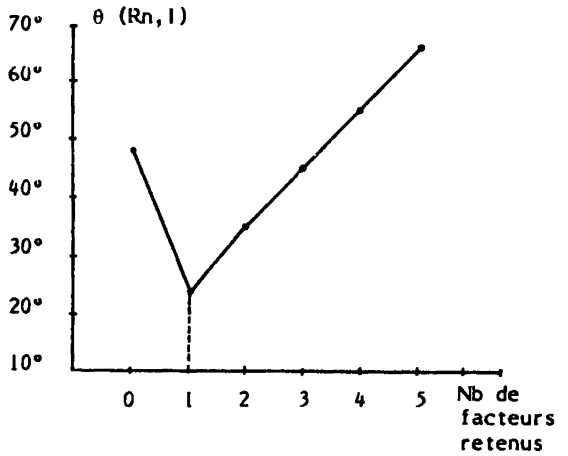
$n=0 \Rightarrow R_n = X'X \cos\theta = \dfrac{q}{\sqrt{q}\ \sqrt{q(q-1)\ p^2+q}}$

$n>0 \Rightarrow R_n = X'X - \hat{Q}_n \cos\theta = \sqrt{\dfrac{q-n}{q}}$

Dans le graphique nous considérons
q = 6

Nb de facteurs retenus :	0	1	2	3	4	5
$\theta\ (R_n, I)$:	18°	24°	35°	45°	55°	66°

pour P=0,5



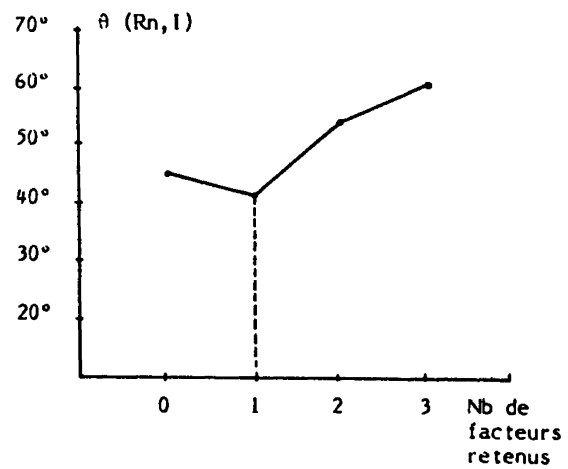
Cas 3 :

La matrice de corrélations $\begin{bmatrix} 1 & 0.7 & 0.6 & 0.4 \\ & 1 & 0.4 & 0.6 \\ & & 1 & 0.7 \\ & & & 1 \end{bmatrix}$ correspond à une situation où les coefficients de corrélation multiple de chacune des q variables, sachant les q-1 autres, sont tous égaux ; les vecteurs propres fournissent les q contrastes orthogonaux du plan d'expérience à 2^q facteurs.

Les valeurs propres fournies par l'ACP sont :

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

Nb de facteurs retenus :	0	1	2	3
$\theta(R_n, I)$:	45°	41°	54°	60°



REFERENCES

(1) ANASTASSAKOS I., D'AUBIGNY G.
"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, 1984, Vol XXXII, n° 2, pp 45-57.