LES CAHIERS DE L'ANALYSE DES DONNÉES

P. CAZÈS

Problème d'analyse des données

Les cahiers de l'analyse des données, tome 1, n° 2 (1976), p. 121-125

<http://www.numdam.org/item?id=CAD_1976__1_2_121_0>

Les Cahiers de l'Analyse des Données
Vol. I - 1976 - n° 2 - p. 121-125

PROBLÈME D'ANALYSE DES DONNÉES

Proposé en DEA de statistique de l'Université
Pierre-et-Marie-Curie en septembre 1975

d'après la solution proposée par P. Cazes (1)

1. *Origine du présent problème :*

Des économistes ont noté un ensemble I de 43 pays suivant un ensemble Q de 15 critères tels que Inflation, Balance des Paiements, Stabilité Politique etc... En combinant linéairement ces notes primaires pondérées par des coefficients de leur choix, les experts ont calculé pour chaque pays 4 notes agrégées : la note d'évaluation générale EVG, et les notes FIN, EXC, POL, (relatives aux finances, à l'exécution des contrats et à la politique (*).

Les notes primaires, étant toutes comprises entre 0 et 4, ont été dédoublées, d'où un tableau I x J (43 pays) x (2 x 15 notes) (où toutes les lignes -pays- ont même total). L'analyse de ce tableau fournit un premier facteur général très fort (τ_1 = 81,7 %); puis un deuxième facteur lié au développement (τ_2 = 6,6 %); et un troisième dominé par la politique (τ_3 = 3,8 %). Les quatre notes agrégées, EVG, FIN, EXC, POL placées en éléments supplémentaires ont un premier facteur nettement prédominant; mais on s'interroge sur le rapport entre la note POL et le troisième facteur.

Pour cela il s'impose de faire des calculs de corrélation, de corrélation résiduelle, etc... On désire simplifier ces calculs en passant par les résultats de l'analyse factorielle. D'où le problème suivant qui commence par des questions de cours et s'achève en une application numérique.

2. *Rappel des notations.*

I, J, S des ensembles finis;

$f_{IJ} = \{f_{ij} | i \in I, j \in J\}$: une loi de probabilité sur I x J;

$f_I = \{f_i | i \in I\}$; $f_i = \Sigma\{f_{ij} | j \in J\}$: la loi marginale sur I;

$k_{IS} = \{k(i,s) | i \in I, s \in S\}$: un tableau de colonnes supplémentaires ajouté à f_{IJ};

$k(s) = \Sigma\{k(i,s) | i \in I\}$: total de la colonne supplémentaire s.

$\varphi_\alpha^I = \{\varphi_\alpha^i | i \in I\}$: un facteur de variance 1 sur I, extrait de f_{IJ}.

---

(1) *ISUP - Laboratoire de Statistique - Université Pierre et Marie Curie-Paris.*

(*) *Cf. Tables à l'usage des investisseurs; d'après M. Greenacre; ce cahier pp. 43-49.*

$F_\alpha^I = \{F_\alpha(i) | i \in I\}$; $G_\alpha^J = \{G_\alpha(j) | j \in J\}$: les facteurs de variance λ_α.

$\{\varphi_\alpha^I | \alpha \in \Lambda\}$ = l'ensemble de tous les facteurs non triviaux issus de f_{IJ}.

3. Enoncé du problème :

Q1 Soient θ^I, ψ^I des fonctions sur l'ensemble I. Donner des formules exprimant la moyenne moy(θ^I) et le coefficient de corrélation corr(θ^I, ψ^I) sur I muni de la loi de probabilité f_I.

Q2 Exprimer en fonction de φ_α^I et des nombres k(i,s) de la colonne supplémentaire s la valeur de $G_\alpha(s)$ (il sera commode d'introduire k(s) dans la formule). Réciproquement, exprimer les k(i,s) en fonction de $\{G_\alpha(s) | \alpha \in A\}$, de k(s), des facteurs $\{\varphi_\alpha^I | \alpha \in A\}$ et de f_I.

Q3 Donner une condition nécessaire et suffisante sur la forme de la colonne des k(i,s) pour que : $\forall\ \alpha \in A : G_\alpha(s) = 0$.

Q4 Notons : pr(i,s) = k(i,s)/(k(s)f_i); donner pour la fonction $pr^{sI} = \{pr(i,s) | i \in I\}$, sa moyenne moy($pr^{sI}$) et ses corrélations avec les facteurs φ_α^I, corr(pr^{sI}, φ_α^I); (on considèrera I muni de la loi f_I; cf Q1).

Q5 On suppose désormais que Card I = n, et que $\forall\ i \in I : f_i = 1/n$. Exprimer corr(k(I,s), k(I,s')) en fonction des $\{G_\alpha(s) | \alpha \in A\}$, $\{G_\alpha(s') | \alpha \in A\}$ (on a noté k(I,s) la colonne supplémentaire s considérée comme fonction sur I).

Q6 Trouver un nombre ρ tel que :
corr((k(I,s) - ρ k(I,s')), k(I,s')) = 0.
La différence k(I,s) - ρ k(I,s') pourra être appelée : colonne s privée de sa composante dans la direction de la colonne s'.

Q7 Exprimer en fonction de ρ et des facteurs $\{G_\alpha(s) | \alpha \in A\}$, $\{G_\alpha(s') | \alpha \in A\}$, le coefficient de corrélation entre k(I,s) - ρ k(I,s') et le facteur φ_β^I (où $\beta \in A$).

Q8 D'après le tableau 1 ci-joint relatif aux éléments supplémentaires de l'analyse qui est à l'origine du présent problème, calculer le coefficient de corrélation entre le troisième facteur et la colonne POL privée de sa composante dans la direction de EVG. (On limitera les calculs aux cinq premiers facteurs. Dans quel sens varierait ce coefficient si on substituait à φ_3^I ce même facteur privé de sa composante dans la direction de EVG.

JSUP	k(s)	α=1	α=2	α=3	α=4	α=5
EVG	990	0.284	-0.020	-0.001	0.005	-0.002
POL	1010	0.283	-0.035	0.052	0.005	0.012
EXC	1007	0.278	-0.003	-0.028	0 019	-0.009
FIN	952	0.312	-0.044	-0.021	0.001	0.015

TABLEAU 1 *masses et coordonnées des éléments supplémentaires*

4. SOLUTION PROPOSEE.

4Q1. Pour toute fonction γ^I définie sur I, On note :

$$moy(\gamma^I) = \Sigma\{f_i\gamma^i | i \in I\};$$

$$var(\gamma^I) = \Sigma\{f_i(\gamma^i - moy(\gamma^I))^2 | i \in I\};$$

et de même pour deux fonctions telles que θ^I et ψ^I :

$$covar(\theta^I,\psi^I) = \Sigma\{f_i(\theta^i - moy(\theta^I))(\psi^i - moy(\psi^I)) | i \in I\};$$

$$corr(\theta^I,\psi^I) = covar(\theta^I,\psi^I) \times (var(\theta^I)\cdot var(\psi^I))^{-1/2}.$$

4Q2. Notons p_I^s le profil de l'élément supplémentaire s :

$$p_I^s = \{p_i^s | i \in I\}; \quad p_i^s = k(i,s)/k(s).$$

p_I^s est une mesure de masse totale 1; la différence $(p_I^s - f_I)$ est donc une mesure de masse totale zéro. Dans l'hyperplan h_I des mesures de masse totale 0 (sur I) on a le système orthonormé des vecteurs axiaux factoriels $e_{\alpha I}$ (mesures ayant pour densité φ_α^I), $\{e_{\alpha I} | \alpha \in A\}$:

$$e_{\alpha I} = \{e_{\alpha i} | i \in I\}; \qquad e_{\alpha i} = \varphi_\alpha^i \cdot f_i;$$

Ce système n'est pas nécessairement complet; il l'est généralement si Card I < Card J. Les $G_\alpha(s)$ sont les coordonnées de $(p_I^s - f_I)$ sur ces axes :

$$G_\alpha(s) = <(p_I^s - f_I), e_{\alpha I}>_{f_I} = \Sigma\{p_i^s e_{\alpha i}/f_i | i \in I\}$$

$$= \Sigma(p_i^s \varphi_\alpha^i | i \in I\} = \Sigma\{k(i,s)/k(s)\ \varphi_\alpha^i | i \in I\};$$

(où l'on a tenu compte de ce que $< f_I, e_{\alpha I} > = 0$). Si le système $e_{\alpha I}$ est complet, on a :

$$p_I^s = f_I + \Sigma\{G_\alpha(s) e_{\alpha I} | \alpha \in A\}$$

$$k(i,s) = k(s)\cdot f_i(1 + \Sigma\{G_\alpha(s)\ \varphi_\alpha^i | \alpha \in A\}$$

4Q3. Sous réserve que le système $\{e_{\alpha I} | \alpha \in A\}$ soit complet, la nullité de tous les $G_\alpha(s)$ équivaut à ce que $p_I^s = f_I$.

4Q4. Il importe ici de se souvenir que la mesure de référence f_I étant fixée une fois pour toutes, à toute fonction θ^I correspond une mesure $(\theta f)_I$ ayant θ pour densité : $(\theta f)_i = \theta^i f_i$; et que $moy(\theta^I)$ = masse totale $(\theta f)_I = \Sigma\{\theta^i f_i | i \in I\}$. Par exemple, la fonction $pr^{Is} = \{pr(i,s) | i \in I\}$ est la densité du profil p_I^s de l'élément supplémentaire s; $moy(pr^{Is})$ = masse $(p_I^s) = 1$. Le coefficient de corrélation entre deux fonctions θ^I et φ^I de moyenne nulle, n'est autre que le cosinus de l'angle formé par les vecteurs $(\theta f)_I$, $(\varphi f)_I$ de I'_I (mesures de masse 0 ayant θ et φ pour densité), par exemple :

$$\begin{aligned}\mathrm{Corr}(pr^{Is}, \varphi_\alpha^I) &= \mathrm{Corr}(pr^{Is} - 1, \varphi_\alpha^I)\\ &= \cos((p_I^s - f_I), e_{\alpha I})\\ &= \langle (p_I^s - f_I), e_{\alpha I}\rangle \ / \|p_I^s - f_I\|\\ &= G_\alpha(s)/\|p_I^s - f_I\|;\end{aligned}$$

si le système orthonormé $\{e_{\alpha I} | \alpha \in A\}$ est complet dans H_I, (cf Q3) on a :

$$\|p_I^s - f_I\|^2 = \Sigma\{G_\alpha(s)^2 | \alpha \in A\}$$

$$\mathrm{corr}(pr^{Is}, \varphi_\alpha^I) = G_\alpha(s)/(\Sigma\{G_\beta(s)^2 | \beta \in A\})^{1/2}$$

4Q5. Si les f_i sont égaux, la fonction k(I,s) et la densité pr^{Is} du profil de s sont proportionnelles entre elles; on doit donc calculer $\mathrm{corr}(pr^{Is}, pr^{Is'})$. D'après les principes rappelés en Q4 il vient :

$$\mathrm{corr}(k(I,s), k(I,s')) = \cos((p_I^s - f_I), (p_I^{s'} - f_I))$$

$$= \langle (p_I^s - f_I), (p_I^{s'} - f_I)\rangle / (\|p_I^s - f_I\| \cdot \|p_I^{s'} - f_I\|).$$

Si le système $\{e_{\alpha I} | \alpha \in A\}$ est complet dans H_I, les produits scalaires et les normes peuvent être calculés dans ce système d'axes orthonormé :

$$\langle (p_I^s - f_I), (p_I^{s'} - f_I)\rangle = \Sigma\{G_\alpha(s)\, G_\alpha(s') | \alpha \in A\};$$

$$\|p_I^s - f_I\|^2 = \Sigma\{G_\alpha(s)^2 | \alpha \in A\};$$

$$\|p_I^{s'} - f_I\|^2 = \Sigma\{G_\alpha(s')^2 | \alpha \in A\}.$$

4Q6. On applique les mêmes principes :

$$\begin{aligned}&\mathrm{corr}((k(I,s) - \rho k(I,s')), k(I,s'))\\ &\quad = \mathrm{corr}((k(s)pr^{Is} - \rho k(s')pr^{Is'}), pr^{Is'})\\ &\quad = \cos((k(s)(p_I^s - f_I) - \rho k(s')(p_I^{s'} - f_I)), (p_I^{s'} - f_I)).\end{aligned}$$

Ce coefficient de corrélation s'annule si :

$$k(s) \langle (p_I^s - f_I), (p_I^{s'} - f_I)\rangle = \rho k(s') \|p_I^{s'} - f_I\|^2;$$

sous l'hypothèse que $\{e_{\alpha I} | \alpha \in A\}$ est complet, cela fait :

$$\rho = (k(s)/k(s'))(\Sigma\{G_\alpha(s)\, G_\alpha(s') | \alpha \in A\}/\Sigma\{G_\alpha(s')^2 | \alpha \in A\}).$$

4Q7. Dans la même voie, on trouve :

$$\begin{aligned}&\mathrm{corr}((k(I,s) - \rho k(I,s')), \varphi_\beta^I)\\ &\quad = \cos((k(s)(p_I^s - f_I) - \rho k(s')(p_I^{s'} - f_I)), e_{\beta I})\\ &\quad = \frac{k(s)G_\beta(s) - \rho k(s')G_\beta(s')}{(\Sigma\{(k(s)G_\alpha(s) - \rho k(s')G_\alpha(s'))^2 | \alpha \in A\})^{1/2}}\end{aligned}$$

(formule où l'on s'est souvenu que $\cos(a_I, b_I) = \langle a_I, b_I\rangle / (\|a_I\| \cdot \|b_I\|)$).

4Q8. L'objet de cette huitième question est de préciser l'interprétation du facteur 3 : il apparait d'abord que ce facteur est fortement lié à l'indice POL qui caractérise (selon les données du Pr. Haner) les conditions politiques; on va tenter d'obtenir entre POL et φ_3 une quasi identité en retranchant leurs composants dans la direction de la note EVG d'évaluation générale.

Notons s = POL; s' = EVG; on a k(s) = 1010; k(s') = 990;

$$\{G_\alpha(s') - G_\alpha(s)\} = \{+.001;\ +.015;\ -.053;\ +.000;\ -.014\}$$

d'où l'on tire (cf Q6), $\rho = 1{,}0194$. Pour calculer $\mathrm{corr}(k(I,s)-\rho k(I,s'),\varphi_3^I)$, calculons (au millième près) :

$$B_\alpha = G_\alpha(s) - \rho(k(s')/k(s))\ G_\alpha(s') = G_\alpha(s) - G_\alpha(s') + 8.10^{-4} G_\alpha(s');$$

$$\{B_\alpha \mid \alpha = 1,\ldots,5\} = \{-.000;\ -.015;\ +.053;\ +.000;\ +.014\}$$

On a d'après Q7 :

$$\mathrm{corr}(k(I,s) - \rho k(I,s')),\varphi_3^I) = B_3/(\Sigma\{B_\alpha^2 \mid \alpha = 1,\ldots,5\})^{1/2} = 0{,}932;$$

c'est bien un coefficient très voisin de 1. Reste à apprécier ce que devient ce coefficient si φ_3 est lui aussi privé de sa composante $\rho' k(I,s')$ dans la direction de EVG. On a :

$$\begin{aligned}
&\mathrm{corr}(k(I,s) - \rho k(I,s'),\ \varphi_3^I - \rho' k(I,s')) \\
&\quad = \cos(((p_I^s - f_I) - \rho(k(s')/k(s))(p_I^{s'} - f_I)),\ e_{3I} - \rho' k(s') p_I^{s'}) \\
&\quad = \cos(u_I,\ e_{3I} - \rho' k(s') p_I^{s'}) \\
&\quad = \langle u_I,\ e_{3I} - \rho' k(s') p_I^{s'} \rangle / (\ \|u_I\| \cdot \|e_{3I} - \rho' k(s') p_I^{s'}\|\ )
\end{aligned}$$

(où l'on a écrit en bref $u_I = (p_I^s - f_I) - \rho(k(s')/k(s))(p_I^{s'} - f_I)$; l'adjonction du terme en ρ' ne modifie pas le produit scalaire du numérateur (ρ a été justement choisi pour que $\langle u_I,\ p_I^{s'} \rangle = 0$); de plus e_{3I} se trouve quelque peu raccourci quand on lui retranche sa composante dans une direction, quelle qu'elle soit (ici celle de EVG); donc la corrélation ne peut être qu'accrue par le terme en ρ'; un calcul numérique précis montre qu'au demeurant cette variation est inférieure à 10^{-4} !