LES CAHIERS DE L'ANALYSE DES DONNÉES

A. EL OUADRANI
PH. NABHAN

Corrélations entre facteurs calculés sur un ensemble d'individus d'après l'analyse d'un sous-tableau de Burt généralisé

Les cahiers de l'analyse des données, tome 19, n° 4 (1994), p. 417-422

<http://www.numdam.org/item?id=CAD_1994__19_4_417_0>

Les Cahiers de l'Analyse des Données
Vol. XIX - 1994 - n°4 - pp.417-422.

CORRÉLATIONS ENTRE FACTEURS CALCULÉS SUR UN ENSEMBLE D'INDIVIDUS D'APRÈS L'ANALYSE D'UN SOUS-TABLEAU DE BURT GÉNÉRALISÉ

[FACT. BURT GÉN]

A. El OUADRANI, Ph. NABHAN

1 Position du problème

Afin d'étudier les corrélations entre deux groupes de variables, l'usage s'est établi de passer par l'analyse d'un sous-tableau rectangulaire d'un tableau de BURT. Nous citerons comme exemples la comparaison entre structure et décor d'un ensemble de lampes, dans [LAMPES SAMOS]; et, dans [BURT. COD. BARY.], l'étude des propriétés mécaniques du bois, en distinguant la contrainte de rupture des propriétés mesurables dans un essai non destructif.

Nous nous proposons ici de démontrer certaines propriétés des différents ensembles de facteurs que l'on considère dans cette méthode de comparaison; en nous plaçant dans un cadre aussi général que possible.

1.1 Cas d'un tableau de BURT usuel

De façon précise, soit un ensemble I d'individus; décrit chacun par un ensemble Q de variables (ou questions d'un dossier), que l'on suppose partagé en deux sous-ensembles Q_a et Q_b; chaque variable q étant découpée en un ensemble J_q de modalités.

On écrit:

$$Q = Q_a \cup Q_b \; ; \; J_a = \cup\{J_q \mid q \in Q_a\} \; ; \; J_b = \cup\{J_q \mid q \in Q_b\} \; ; \; J = J_a \cup J_b \; ;$$

dans le tableau de BURT, $J \times J$, associé à un tel codage, on distingue le sous-rectangle de BURT, $J_a \times J_b$. À l'analyse de ce sous-rectangle, l'ensemble I des individus peut être adjoint en supplément de deux manières: soit comme ensemble de lignes, soit comme ensemble de colonnes; d'où, pour I, sur l'espace engendré par les axes factoriels, deux représentations qu'il est utile de comparer.

Les cahiers de l'analyse des données 0339-3097/94/04/06/$ 4.00/© Gauthier-Villars

Pratiquement, le calcul s'effectue en soumettant à l'analyse le tableau de BURT J × J tout entier; mais en ne gardant en principal que l'ensemble J_a de lignes et l'ensemble J_b de colonnes.

Pour le calcul, il n'est pas nécessaire de supposer que les deux sous-ensembles, Q_a et Q_b, sont d'intersection vide; cette condition sera observée, toutefois, si l'on veut éviter que la redondance entre les deux blocs ne domine, dans les résultats de l'analyse, la corrélation entre groupes de variables rendant compte d'aspects différents des individus.

1.2 Généralisation du sous-tableau de BURT

Des deux exemples cités ci-dessus, l'un (l'étude du bois) ne concerne pas un tableau de BURT, *stricto sensu*, mais une généralisation au cas d'un codage barycentrique: c'est pourquoi nous avons dit qu'il convenait de prendre, pour l'énoncé des résultats et leur démonstration, le cadre le plus général où nous sachions comment procéder.

On part donc de deux tableaux $k_a(I, J_b)$ et $k_b(I, J_a)$ satisfaisant à la condition de proportionalité suivante:

$$\forall\, i \in I : k_a(i, J_b)/k_a = k_b(i, J_a)/k_b \quad ;$$

les quantités introduites étant définies ci-dessous par des sommes (où j_a désigne un élément arbitraire de J_a):

$$k_a(i, J_b) = \Sigma\{k_a(i, j_b) \mid j_b \in J_b\} \quad ;$$

$$k_a = \Sigma\{k_a(i, J_b) \mid i \in I\} = \Sigma\{k_a(i, j_b) \mid i \in I\,;\, j_b \in J_b\} \;;$$

et de même pour $k_b(i, J_a)$ et k_b, en remplaçant a par b. On peut noter:

$$k(i) = k_a(i, J_b)/k_a = k_b(i, J_a)/k_b \quad ; \quad \Sigma\{k(i) \mid i \in I\} = 1 \quad ;$$

où k(i) tient le rôle d'une fréquence.

Comme l'analogue du sous-rectangle de BURT considéré d'abord, on définit le tableau rectangulaire $K(J_a \times J_b)$:

$$K(j_a, j_b) = \Sigma\{k_b(i, j_a).k_a(i, j_b) / (k_a.k_b.k(i)) \mid i \in I\} \;;$$

les marges de K sur J_a et J_b sont respectivement données par:

$$K(j_a) = \Sigma\{k_b(i, j_a).k_a(i, j_b) / (k_a.k_b.k(i)) \mid i \in I\,;\, j_b \in J_b\} = k_b(j_a)/k_b \;;$$

et de même $K(j_b) = k_a(j_b)/k_a$.

Dans le tableau K, nous conviendrons de dire que J_a et J_b sont,

respectivement, l'ensemble des lignes et l'ensemble des colonnes. Au tableau $J_a \times J_b$, l'ensemble I peut être adjoint en supplément de deux manières, comme un ensemble I_a de lignes, par le tableau $k_a(I, J_b)$; et, par le tableau $k_b(I, J_a)$, comme un ensemble I_b de colonnes.

N.B. On notera que chacun des tableaux k_a et k_b a respectivement pour ensemble de variables celui des deux ensembles J_b et J_a dont l'indice diffère du sien: en effet, par le tableau k_a, un individu i est adjoint comme ligne supplémentaire au tableau K, et donc ajouté à l'ensemble J_a des lignes de l'analyse principale; et, de même, k_b définit des colonnes supplémentaires ajoutées à J_b.

Ces définitions étant posées en toute généralité, nous supposerons désormais que l'on a la condition de normalisation:

$$k_a = k_b = 1 \ ; \quad \text{d'où:} \quad \forall\, i \in I \ : \ k_a(i) = k_b(i) = k(i) \quad ;$$

ce qui n'influe pas sur le résultat des calculs de facteurs, tout en simplifiant l'écriture des démonstrations que nous avons en vue.

2 Propriétés des facteurs associés à un sous-tableau de BURT

2.1 Formules de définition des deux groupes de facteurs sur I

Un couple de facteurs associés issu du tableau K sera noté $(F_\gamma(J_a), G_\gamma(J_b))$; et de même, pour les facteurs sur I, respectivement $F_\gamma(I_a)$ et $G_\gamma(I_b)$; avec, entre ces facteurs, tous de moyenne nulle, les relations:

$$G_\gamma(i) = (1/\sqrt{(\lambda_\gamma)}) \cdot \Sigma\{(k_b(i, j_a) / k(i)) \cdot F_\gamma(j_a) \mid j_a \in J_a\} \quad ;$$

$$F_\gamma(i) = (1/\sqrt{(\lambda_\gamma)}) \cdot \Sigma\{(k_a(i, j_b) / k(i)) \cdot G_\gamma(j_b) \mid j_b \in J_b\} \quad ;$$

$$\begin{aligned} G_\gamma(j_b) &= (1/\sqrt{(\lambda_\gamma)}) \cdot \Sigma\{(K(j_a, j_b) / K(j_b)) \cdot F_\gamma(j_a) \mid j_a \in J_a\} \\ &= (1/\sqrt{(\lambda_\gamma)}) \cdot \\ &\qquad \Sigma\{(k_a(i, j_b) \cdot k_b(i, j_a) / (k(i).k_a(j_b)) \cdot F_\gamma(j_a) \mid j_a \in J_a \ ; \ i \in I\} \ ; \\ &= \Sigma\{(k_a(i, j_b) / k_a(j_b)) \cdot G_\gamma(i) \mid i \in I\} \quad ; \end{aligned}$$

$$\begin{aligned} F_\gamma(j_a) &= (1/\sqrt{(\lambda_\gamma)}) \cdot \Sigma\{(K(j_a, j_b) / K(j_a)) \cdot G_\gamma(j_b) \mid j_b \in J_b\} \\ &= (1/\sqrt{(\lambda_\gamma)}) \cdot \\ &\qquad \Sigma\{(k_b(i, j_a) \cdot k_a(i, j_b) / (k(i).k_b(j_a)) \cdot G_\gamma(j_b) \mid j_b \in J_b \ ; \ i \in I\} \ ; \\ &= \Sigma\{(k_b(i, j_a) / k_b(j_a)) \cdot F_\gamma(i) \mid i \in I\} \quad ; \end{aligned}$$

2.2 Bornes pour la variance des facteurs sur I

Ainsi, il apparaît que l'on passe de $F_\gamma(I_a)$ à $F_\gamma(J_a)$ par la transition probabiliste associée au tableau k_b; et de même pour les G_γ et k_a. De ce qu'une transition diminue la variance d'une fonction, il résulte que:

$$\mathrm{Var}(F_\gamma(J_a)) \leq \mathrm{Var}(F_\gamma(I)) \quad ; \quad \mathrm{Var}(G_\gamma(J_b)) \leq \mathrm{Var}(G_\gamma(I)) \quad ;$$

mais, d'autre part, les facteurs sur I sont définis par transition à partir de ceux sur J_a, au coefficient $(1/\sqrt{(\lambda_\gamma)})$ près; d'où:

$$\mathrm{Var}(F_\gamma(I)) \leq (1/\lambda_\gamma).\mathrm{Var}(F_\gamma(J_a)) \quad ; \quad \mathrm{Var}(G_\gamma(I)) \leq (1/\lambda_\gamma).\mathrm{Var}(G_\gamma(J_b)) \quad ;$$

et, puisque les λ_γ sont ≤ 1, on a les inégalités compatibles:

$$\mathrm{Var}(F_\gamma(J_a)) \leq \mathrm{Var}(F_\gamma(I)) \leq (1/.\lambda_\gamma).\mathrm{Var}(F_\gamma(J_a)) \quad ;$$

$$\mathrm{Var}(G_\gamma(J_b)) \leq \mathrm{Var}(G_\gamma(I)) \leq (1/\lambda_\gamma).\mathrm{Var}(G_\gamma(J_b)) \quad ;$$

$$\lambda_\gamma \leq \mathrm{Var}(F_\gamma(I)) \leq 1 \quad ; \quad \lambda_\gamma \leq \mathrm{Var}(G_\gamma(I)) \leq 1 \ ;$$

dans le cas, très particulier, où l'analyse porte sur un tableau de BURT usuel et où, de plus, $J_a = J_b = J$, tout entier, les facteurs $F_\gamma(I)$ et $G_\gamma(I)$ de même rang coïncident, avec pour variance $\sqrt{(\lambda_\gamma)}$; i.e. la moyenne géométrique des bornes trouvées ci-dessus pour encadrer $\mathrm{Var}(F_\gamma(I))$ et $\mathrm{Var}(G_\gamma(I))$.

2.3 Covariance entre les deux groupes de facteurs sur I

Le calcul des variances et covariances des facteurs sur I, servira à établir, entre le système des $\{F_\gamma(I) \mid \gamma \in \Gamma\}$ et le système des $\{G_\gamma(I) \mid \gamma \in \Gamma\}$, une relation qui généralise celle des trièdres supplémentaires.

De façon précise, on a pour deux facteurs ayant respectivement pour indice γ et ξ, distincts ou non :

$$\Sigma\{F_\gamma(i) \,.\, G_\xi(i) \,.\, k(i) \mid i \in I\} = \mathrm{Cov}((F_\gamma(I), G_\xi(I)) =$$

$$(1/\sqrt{(\lambda_\gamma.\lambda_\xi)}) \,.$$

$$\Sigma\{(k_b(i, j_a) \,.\, k_a(i, j_b) / k(i)) \,.\, F_\gamma(j_a).\, G_\xi(j_b) \mid j_a \in J_a \,;\, j_b \in J_b;\ i \in I\}$$

$$= (1/\sqrt{(\lambda_\gamma.\lambda_\xi)}).\Sigma\{K(j_a, j_b) \,.\, F_\gamma(j_a) \,.\, G_\xi(j_b) \mid j_a \in J_a \,;\, j_b \in J_b\}$$

$$= \mathrm{delt}(\gamma, \xi) \,.\, \sqrt{(\lambda_\gamma)} \quad ;$$

où on a noté comme une fonction de deux variables, delt(γ, ξ), le tenseur δ_γ^ξ usuel, qui vaut 1 si $\gamma = \xi$, et zéro sinon.

3.4 Corrélations entre les deux groupes de facteurs sur I et supplémentarité

Quant aux corrélations entre facteurs sur I, on a d'abord l'égalité:

$$(\gamma \neq \xi) \Rightarrow \text{corr}(F_\gamma(I), G_\xi(I)) = 0 \ ;$$

d'autre part, de la formule classique :

$$\text{corr}(F_\gamma(I), G_\gamma(I)) = \text{Cov}(F_\gamma(I), G_\gamma(I)) / \sqrt{(\text{Var}(F_\gamma(I)).\text{Var}(G_\gamma(I)))} \ ;$$

de l'égalité: $\text{Cov}(F_\gamma(I), G_\gamma(I)) = \sqrt{(\lambda_\gamma)}$, et des majorations trouvées ci-dessus pour les variances de $F_\gamma(I)$ et $G_\gamma(I)$, on déduit l'inégalité:

$$\sqrt{(\lambda_\gamma)} \leq \text{corr}(F_\gamma(I), G_\gamma(I)) \leq 1 \ .$$

Plaçons-nous dans l'espace R^I des fonctions sur I, avec la norme euclidienne associée aux pondérations k(i) ; afin de retrouver le langage géométrique usuel, désignons respectivement par aXe_γ et bXe_γ les vecteurs unitaires (ou fonctions de variance 1) proportionnels, respectivement, aux facteurs $F_\gamma(I)$ et $G_\gamma(I)$:

on a deux systèmes de vecteurs:

$$\{aXe_\gamma \mid \gamma \in \Gamma\} \ ; \ \{bXe_\gamma \mid \gamma \in \Gamma\} \ ;$$

hormis le cas où $k_a(I, J_b) = k_b(I, J_a)$, ni l'un ni l'autre des systèmes n'est généralement orthonormé; mais chacun des vecteurs de l'un des systèmes est orthogonal à tout vecteur de l'autre système ayant un indice différent du sien; on peut encore écrire, en notant $\mathbf{L}\{V\}$le sous-espace engendré par un ensemble V de vecteurs:

$$\forall \ \gamma \in \Gamma : \ aXe_\gamma \ \text{orthogonal à} \ \mathbf{L}\{bXe_\xi \mid \xi \in \Gamma ; \xi \neq \gamma\} \ ;$$

$$\forall \ \gamma \in \Gamma : \ bXe_\gamma \ \text{orthogonal à} \ \mathbf{L}\{aXe_\xi \mid \xi \in \Gamma ; \xi \neq \gamma\} \ ;$$

il faut toutefois noter qu'à la différence de ce qui est pour les trièdres supplémentaires en dimension 3, chacun des deux systèmes de vecteurs n'est pas déterminé par l'autre; parce que cardΓ est, en général, inférieur à la dimension, cardI, de R^I (ou même à cardI-1, dimension de l'espace des fonctions de moyenne nulle).

3 Note sur la validité des corrélations trouvées entre les deux groupes de facteurs sur I

En analyse des correspondances, l'appréciation de la validité des résultats repose sur l'unterprétation même: dans la mesure où, e.g., ne peut être issue du hasard la disposition régulière des suites de modalités d'un groupe de variables découpées en classes.

Mais, dans le cas présent, il semble facile de recourir à la simulation; à condition que tous les individus i aient même masse k(i)=1/cardI. Dans ce cas, le tableau $k_a(I, J_b)$ étant conservé tel quel, on soumettra l'ensemble, I, des lignes du tableau $k_b(I, J_a)$, à des permutations aléatoires σ; d'où pour chaque σ, une analyse factorielle du tableau K correspondant; et des valeurs propres et corrélations entre facteurs; à la distribution desquels on comparera le résultat obtenu pour les données initiales, σ = identité. Si, comme on peut le présumer, la corrélation entre données naturelles est forte relativement à celle qui se rencontre dans les données simulées; la validité des conclusions tirées de l'analyse sera confirmée avec un nombre de simulations assez faible; (même si quelque 200 simulations seraient nécessaires pour calculer des seuils classiques à 5% ou 2%).

Références bibliographiques

A. El OUADRANI : "Généralisation du tableau de BURT et de l'analyse de ses sous-tableaux dans le cas d'un codage barycentrique", [BURT COD. BARY.], in *CAD*, Vol.XIX, n°2, pp. 229-246; (1994).

A. ATHANASSIADIS, N. POULOU-PAPADIMITRIOU :"Les lampes paléo-chrétiennes de l'île de Samos: examen statistique", [LAMPES SAMOS], in *CAD*, Vol.XIX, n°3, pp. 305-322; (1994).