

REVUE DE STATISTIQUE APPLIQUÉE

A. LAZRAQ

R. CLÉROUX

Tests d'homogénéité entre indices de redondance pour des lois elliptiques

Revue de statistique appliquée, tome 40, n° 3 (1992), p. 19-33

http://www.numdam.org/item?id=RSA_1992__40_3_19_0

© Société française de statistique, 1992, tous droits réservés.

L'accès aux archives de la revue « Revue de statistique appliquée » (<http://www.sfds.asso.fr/publicat/rsa.htm>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
<http://www.numdam.org/>

TESTS D'HOMOGÉNÉITÉ ENTRE INDICES DE REDONDANCE POUR DES LOIS ELLIPTIQUES

A. Lazraq⁽¹⁾ et R. Cléroux⁽²⁾

(1) Ecole Nationale de l'Industrie Minérale, Rabat, Maroc

(2) Département d'informatique et de recherche opérationnelle,
Université de Montréal, Canada

RÉSUMÉ

Dans cet article on s'intéresse à l'indice de redondance entre deux ensembles de variables. On obtient des résultats distributionnels dans le cas où la loi sous-jacente se trouve dans la classe des lois elliptiques. On peut alors tester l'homogénéité entre plusieurs indices de redondance ainsi que la nullité de plusieurs indices de redondance. La loi conjointe asymptotique de plusieurs indices de redondance peut également être obtenue afin de tester l'homogénéité dans le cas où les populations considérées ne sont pas indépendantes.

Mots-clés : Association multivariée, Corrélation vectorielle, Mesures de redondances, Lois elliptiques, Tests d'homogénéité.

ABSTRACT

Some measures of multivariate association between two sets of variables are considered. Some distributional results are obtained, when the underlying distribution is in the class of elliptical distributions, which allow testing the hypothesis of homogeneity between a set of such measures of association, and the hypothesis that several such measures are zero. The joint asymptotic distribution of such measures of association can also be obtained to test the hypothesis of homogeneity when the populations involved are not independent.

Key-words : Multivariate association, Vector correlation, Redundancy measurements, Elliptical distribution, Test of homogeneity.

Introduction

Dans cette article nous nous intéressons aux mesures d'association entre deux ensembles de variables. Plusieurs auteurs ont considéré des mesures de corrélation,

redondance ou association entre deux vecteurs aléatoires, deux ensembles de variables ou deux matrices de données : voir par exemple Hotelling (1936), Masuyama (1939, 1941), Roseboom (1965), Stewart et Love (1968), Kshirsagar (1969), Escoufier (1973), Coxhead (1974), Cramer (1974), Lingoes et Shonemann (1974), Shaffer et Gillo (1974), Gleason (1976), Robert et Escoufier (1976), Cramer et Nicewander (1979), Stephens (1979) et finalement Ramsay, ten Berge et Styan (1984). Ces mesures furent groupées en deux classes par Cramer et Nicewander (1979) : mesures de redondance qui visent à prédire un ensemble de variables par un autre et mesures d'association ou de corrélation vectorielle qui généralisent le concept de coefficient de corrélation à deux ensembles de variables.

On rappelle d'abord l'indice de redondance de Stewart et Love (1968) et Gleason (1976) (Section 1) et on obtient des résultats distributionnels dans le cas d'une loi elliptique (qui généralise la loi normale multivariée) en vue de tester l'homogénéité d'un ensemble de tels indices. Les Sections 2 et 3 sont consacrées aux lois asymptotiques de l'indice de redondance sous différentes hypothèses. Dans la Section 4 on construit un test d'homogénéité entre plusieurs indices de redondance. Dans la Section 5 on teste la nullité de plusieurs indices. Finalement, dans la Section 6, on obtient la distribution conjointe asymptotique de plusieurs indices et on teste leur égalité dans le cas où les populations considérées ne sont pas indépendantes. Nous traitons plusieurs exemples.

1. Un indice de redondance

Soient $X^{(1)} : p \times 1$ et $X^{(2)} : q \times 1$ deux vecteurs aléatoires avec moyennes $E(X^{(i)}) = \mu^{(i)}$ et matrices de covariances $\Sigma_{ij} = E(X^{(i)} - \mu^{(i)})(X^{(j)} - \mu^{(j)})'$ pour $i = 1, 2$ et $j = 1, 2$. Ecrivons

$$X = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu^{(1)} \\ \mu^{(2)} \end{pmatrix} \text{ et } \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{21} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

et supposons Σ définie positive. Les paramètres μ et Σ sont habituellement inconnus et sont estimés par $\bar{X}$ et S définis ci-dessous. Soit $\begin{pmatrix} X_{\alpha}^{(1)} \\ X_{\alpha}^{(2)} \end{pmatrix}$, $\alpha =$

$1, 2, \dots, n$, un échantillon aléatoire de X et soient $\bar{X}^{(i)} = \frac{1}{n} \sum_{\alpha=1}^n X_{\alpha}^{(i)}$ et $S_{ij} =$

$\frac{1}{n-1} \sum_{\alpha=1}^n (X_{\alpha}^{(i)} - \bar{X}^{(i)})(X_{\alpha}^{(j)} - \bar{X}^{(j)})'$ pour $i = 1, 2$ et $j = 1, 2$. Finalement,

soient $\bar{X} = \begin{pmatrix} \bar{X}^{(1)} \\ \bar{X}^{(2)} \end{pmatrix}$ et $S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}$.

L'indice de redondance s'inscrit dans un contexte de prédiction de $X^{(1)}$ par $X^{(2)}$. Le modèle de régression linéaire multivariée s'écrit

$$X_1 = X_2 B + E \tag{1}$$

où X_1 est une matrice $n \times p$ de p variables mesurées sur n individus, X_2 est une matrice $n \times q$ de q variables mesurées sur les mêmes individus, B est la matrice $q \times p$ de régression et E est une matrice aléatoire $n \times p$ de résidus. On suppose que les moyennes des colonnes de X_1 et X_2 sont nulles, que les deux matrices sont de plein rang et que $p \leq q$. Alors les matrices de covariance s'écrivent $S_{11} = \frac{1}{n-1} X_1' X_1$, $S_{22} = \frac{1}{n-1} X_2' X_2$, $S_{12} = \frac{1}{n-1} X_1' X_2$ et minimisant $tr(E'E)$, où $tr(\cdot)$ désigne la trace, on obtient $\hat{B} = S_{22}^{-1} S_{21}$ et $\hat{X}_1 = \hat{X}_2 \hat{B}$. Si on écrit $S_{11}^* = \hat{X}_1' \hat{X}_1 = (X_2 \hat{B})' (X_2 \hat{B})$ alors $S_{11}^* = S_{12} S_{22}^{-1} S_{21}$. Finalement, si $SS_{1(i)}$ est le i^e élément de la diagonale de S_{11} et $SS_{1(i)}^*$ celui de S_{11}^* , alors le carré du coefficient de corrélation multiple pour prédire la i^e composante $X_i^{(1)}$ de $X^{(1)}$ par $X^{(2)}$, pour $i = 1, 2, \dots, p$, est donné par

$$R_{1(i) \cdot x_1^{(2)}, x_q^{(2)}}^2 = \frac{SS_{1(i)}^*}{SS_{1(i)}}. \quad (2)$$

Si $\hat{\rho}_1^2, \hat{\rho}_2^2, \dots, \hat{\rho}_p^2$ sont les carrés des corrélations canoniques entre $X^{(1)}$ et $X^{(2)}$ et $\Lambda = \text{diag}(\hat{\rho}_1^2, \hat{\rho}_2^2, \dots, \hat{\rho}_p^2)$, si U est la matrice $p \times p$ telle que la i^e colonne est le vecteur canonique correspondant à $\hat{\rho}_i$ et si $D = U^{-1}$, alors on voit facilement que $S_{11} = D'D$ et $S_{11}^* = D'\Lambda D$.

L'indice de redondance, introduit par Stewart et Love (1968) et généralisé par Gleason (1976), est défini par

$$RI = \frac{tr S_{11}^*}{tr S_{11}} = \frac{tr(S_{12} S_{22}^{-1} S_{21})}{tr S_{11}} = \frac{tr(D'\Lambda D)}{tr(D'D)}. \quad (3)$$

C'est la proportion de la variance de X_1 expliquée par X_2 . Il peut également s'écrire sous la forme

$$RI = \frac{\sum_{i=1}^p \sigma_{1(i)}^2 R_{1(i) \cdot x_1^{(2)}, x_q^{(2)}}^2}{\sum_{i=1}^p \sigma_{1(i)}^2} \quad (4)$$

où $\sigma_{1(i)}^2$ est la variance de la i^e composante de $X^{(1)}$. L'indice de redondance est donc une moyenne pondérée des carrés des coefficients de corrélation multiples entre les composantes de $X^{(1)}$ et le vecteur $X^{(2)}$. Il fut utilisé dans une procédure de régression linéaire multivariée pas à pas par Lazraq et Cléroux (1988). On peut voir que $0 \leq RI \leq 1$, que RI devient égal à r^2 , le carré du coefficient de corrélation simple quand $p = q = 1$ et qu'il se réduit au carré du coefficient de corrélation multiple si $p = 1$ et $q > 1$. Il n'est cependant pas invariant sous les transformations des variables en général. Finalement, au niveau de la population, l'indice de redondance est noté

$$\rho I = \frac{tr(\Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21})}{tr \Sigma_{11}}. \quad (5)$$

2. Loi Asymptotique de RI quand $0 < \rho I < 1$

Posons $m = p + q$. On obtient maintenant la loi asymptotique de RI quand $0 < \rho I < 1$ et quand la loi de X se situe dans la classe des lois elliptiques $E_m(\mu, V)$ ayant pour vecteur moyenne μ , pour matrice de covariance $\Sigma = \alpha V$ où α est un scalaire, et avec paramètre d'aplatissement k (voir Annexe). Rappelons que la loi normale multivariée est une loi elliptique avec $k = 0$.

Pour toute matrice $M : r \times s$, soit $\text{vec}(M)$ le vecteur $rs \times 1$ formé en empilant les colonnes M . On a alors les lemmes et le théorème suivants :

Lemme 1. – Soit $S = (s_{ij})$ une matrice de covariance empirique obtenue à partir d'un échantillon de taille n d'une loi elliptique avec matrice de covariance $\Sigma = (\sigma_{ij}) : m \times m$ et paramètre d'aplatissement k . Alors, lorsque $n \rightarrow \infty$, la loi asymptotique de $n^{\frac{1}{2}}(\text{vec}(S) - \text{vec}(\Sigma))$ est $N(0, \Gamma)$ où $\Gamma = (1+k)(I+K)(\Sigma \otimes \Sigma) + k \text{vec}(\Sigma)(\text{vec}(\Sigma))'$, I est la matrice identité $m^2 \times m^2$, $\otimes$ est le produit matriciel de Kronecker et K est la matrice de commutation définie par $K = \sum_{i=1}^m \sum_{j=1}^m (H_{ij} \otimes H'_{ij})$ où H_{ij} est la matrice $m \times m$ ayant $h_{ij} = 1$ et tous les autres éléments nuls. Les éléments de Γ sont donnés par

$$\text{cov}(s_{ij}, s_{lr}) = (n-1)^{-1} \left[k \sigma_{ij} \sigma_{rl} + (1+k)(\sigma_{ir} \sigma_{jl} + \sigma_{il} \sigma_{jr}) \right]$$

La preuve de ce lemme se trouve dans Muirhead (1982), p. 113.

Lemme 2. – Si $S = (s_{ij}) : m \times m$ est une matrice de covariance empirique et $f(S)$ une fonction réelle continue ayant ses dérivées de premiers et seconds ordres continues par rapport à s_{ij} pour tout (i, j) , si de plus $f(S)$ n'est pas une fonction explicite de la taille n de l'échantillon, alors, lorsque $n \rightarrow \infty$, la loi asymptotique de $n^{\frac{1}{2}}(f(S) - f(\Sigma))$ est $N(0, \sigma^2)$ où $\sigma^2 = n \sum_{i=1}^m \sum_{j=1}^m \sum_{k=1}^m \sum_{l=1}^m \frac{\partial f(\Sigma)}{\partial s_{ij}} \frac{\partial f(\Sigma)}{\partial s_{kl}} \text{cov}(s_{ij}, s_{kl})$, Σ est la matrice de covariance de la population et $\frac{\partial f(\Sigma)}{\partial s_{ij}}$ est la dérivée $\frac{\partial f(S)}{\partial s_{ij}}$ évaluée à $S = \Sigma$.

Ce lemme est prouvé dans Press (1972), p. 94.

Théorème 1. – Soit $s = (s_{ij})$ une matrice de covariance empirique obtenue à partir d'un échantillon de taille n d'une loi elliptique avec matrice de covariance $\Sigma = (\sigma_{ij}) : m \times m$ et paramètre d'aplatissement k . Alors, si $0 < \rho I < 1$, la loi asymptotique, lorsque $n \rightarrow \infty$, de $(n-1)^{\frac{1}{2}}(RI - \rho I)$ est $N(0, 2(\rho I)^2(1+k)\sigma^2)$ où σ^2 est donné par

$$\sigma^2 = \frac{\text{tr}(\Sigma_{11}^2)}{(\text{tr}(\Sigma_{11}))^2} - \frac{4 \text{tr}(\Sigma_{11} \Sigma_{11}^*) - 2 \text{tr}(\Sigma_{11}^2)}{\text{tr}(\Sigma_{11}) \text{tr}(\Sigma_{11}^*)} + \frac{2 \text{tr}(\Sigma_{11} \Sigma_{11}^*) - \text{tr}(\Sigma_{11}^{2*})}{(\text{tr} \Sigma_{11}^*)^2} \quad (6)$$

et où $\Sigma_{11}^* = \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

La preuve de ce théorème est longue et pénible et conséquemment, ne sera pas explicitée. On l'obtient en utilisant les lemmes 1 et 2 ainsi que plusieurs résultats de Dwyer (1967) sur la dérivation de matrice. Ce résultat indique que la variance asymptotique de RI peut augmenter considérablement si la loi sous-jacente n'est pas multivariée. Par exemple, dans le cas de la loi binormale contaminée $.9N(0, \Sigma) + .1N(0, 9\Sigma)$, qui est elliptique avec $k = 1.78$, la variance asymptotique de RI est environ trois fois plus grande que celle correspondant à la loi $N(0, \Sigma)$. On réalise donc que la loi de RI n'est pas robuste à l'absence de la normalité. Notons, en passant, que la loi du coefficient de corrélation simple n'est pas, non plus, robuste à la non-normalité.

3. Loi asymptotique de RI quand $\rho I = 0$

Nous démontrons le théorème suivant :

Théorème 2. – Soit $S = (s_{ij})$ une matrice de covariance empirique obtenue à partir d'un échantillon de taille n d'une loi elliptique avec matrice de covariance $\Sigma = (\sigma_{ij}) : m \times m$ et paramètre d'aplatissement k . Si $\rho I = 0$, la loi asymptotique, lorsque $n \rightarrow \infty$, de nRI est la loi de $(1 + k)(tr(\Sigma_{11}))^{-1} \sum_{i=1}^p \delta_i \sum_{j=1}^q U_{ij}^2$ où $\delta_1 \geq \delta_2 \geq \dots \geq \delta_p$ sont les valeurs propres de Σ_{11} chacune étant prise q fois et où les U_{ij} sont des variables aléatoires iid $N(0, 1)$.

Preuve : On a $RI = \frac{tr(S_{12}S_{22}^{-1}S_{21})}{tr(S_{11})}$ et l'on sait que $S_{11} \xrightarrow{p} \Sigma_{11}$ et $S_{22} \xrightarrow{p} \Sigma_{22}$ où $\xrightarrow{p}$ signifie convergence en probabilité lorsque $n \rightarrow \infty$.

Soit $S_{21} \xrightarrow{l} W$ où $\xrightarrow{l}$ signifie convergence en loi lorsque $n \rightarrow \infty$.

Par le Théorème 4.4 de Billingsley (1968) on a $(S_{21}, S_{22}) \xrightarrow{l} (W, \Sigma_{22})$. Soit $Q(W, \Sigma_{22}) = tr(W'\Sigma_{22}^{-1}W)$. Alors Q est continu en chaque élément de W et de Σ_{22} et par le Théorème 5.1 de Billingsley (1968) on peut écrire $Q(S_{21}, S_{22}) \xrightarrow{l} Q(W, \Sigma_{22})$. Donc $tr(S_{12}S_{22}^{-1}S_{21})$ a la même loi asymptotique que $tr(S_{12}\Sigma_{22}^{-1}S_{21})$ et RI a la même loi asymptotique que $RI^* = \frac{tr(S_{12}\Sigma_{22}^{-1}S_{21})}{tr(\Sigma_{11})}$.

Considérons maintenant le numérateur de RI^* ou plus précisément, considérons

$$n tr(S_{12}\Sigma_{22}^{-1}S_{21}) = n(\text{vec}(S_{21}))'(I_p \otimes \Sigma_{22}^{-1})(\text{vec}(S_{21})). \quad (7)$$

Sous les hypothèses du théorème et quand $\rho I = 0$ on a, lorsque $n \rightarrow \infty$,

$$n^{\frac{1}{2}} \text{vec}(S_{21}) = n^{\frac{1}{2}} B(\text{vec}(S)) \xrightarrow{l} N(0, B\Gamma B')$$

où Γ est définie au Lemme 1, $B = B_1 \otimes B_2$, $B_1 = (I_p \ 0) : p \times m$ et $B_2 = (0 \ I_q) : q \times m$. En conséquence, la forme quadratique (7) est distribuée asymptotiquement comme $\sum_i \omega_i Z_i^2$ où les Z_i sont iid $N(0, 1)$ et où les ω_i sont les valeurs propres de $(B\Lambda B')(I_p \otimes \Sigma_{22}^{-1})$.

Mais lorsque $\rho I = 0$ ou, de façon équivalente, lorsque $\Sigma_{12} = 0$, on a $B(\Sigma \otimes \Sigma)B' = \Sigma_{11} \otimes \Sigma_{22}$, $BK(\Sigma \otimes \Sigma)B' = 0$ où K est la matrice de commutation du Lemme 1, et $B \text{vec}(\Sigma)(\text{vec}(\Sigma))'B' = \text{vec}(\Sigma_{21})(\text{vec}(\Sigma_{21}))' = 0$ de sorte que $(B\Gamma B')(I_p \otimes \Sigma_{22}^{-1}) = (1 + k)(\Sigma_{11} \otimes \Sigma_{22})(I_p \otimes \Sigma_{22}^{-1}) = (1 + k)(\Sigma_{11} I_p) \otimes (\Sigma_{22} \Sigma_{22}^{-1}) = (1 + k)\Sigma_{11} \otimes I_q$.

Donc les ω_i sont les valeurs propres de $(1 + k)\Sigma_{11} \otimes I_q$ c'est-à-dire $(1 + k)\delta_i$, $i = 1, 2, \dots, p$, chaque δ_i étant de multiplicité q , et où $\delta_1, \delta_2, \dots, \delta_p$ sont les valeurs propres de Σ_{11} .

La preuve du théorème est complète.

Les résultats de cette section nous permettent d'effectuer un test asymptotique de $H_0 : \rho I = 0$ contre $H_1 : \rho I > 0$. Le test est le suivant : rejeter H_0 au niveau α si et seulement si $nRI > c_\alpha$ où c_α est le $100(1 - \alpha)^e$ centile de la loi de $(1 + k)(\text{tr}(\Sigma_{11}))^{-1} \sum_{i=1}^p \sum_{j=1}^q \delta_i U_{ij}^2$. Plusieurs algorithmes sont disponibles pour calculer c_α . L'algorithme de Imhof (1961), dont un code informatique se trouve dans Koerts et Abrahamse (1969), peut-être considéré comme exact et c'est celui que nous recommandons. En pratique, les paramètres de la loi sont inconnus et doivent être remplacés par des estimateurs convergents. Les résultats de la Section 2 permettent de calculer la puissance du test précédent.

4. Loi asymptotique de RI sous une suite d'hypothèses alternatives

Considérons la suite d'hypothèses alternatives $H_{1n} : \Sigma_{21} = \frac{A}{\sqrt{n}}$ où A est une matrice $p \times q$ fixée. On obtient la loi asymptotique de nRI sous cette suite d'hypothèses alternatives. On utilise le lemme 5.1 de Cléroux et Ducharme (1989).

Théorème 3. – Soit $S = (s_{ij})$ une matrice de covariance empirique obtenue à partir d'un échantillon de taille n d'une loi elliptique avec matrice de covariance $\Sigma = (\alpha_{ij}) : m \times m$ et paramètre d'aplatissement k . Sous H_{1n} , la loi asymptotique, lorsque $n \rightarrow \infty$, de nRI est la loi de

$$\frac{1 + k}{\text{tr} \Sigma_{11}} \sum_{i=1}^p \sum_{j=1}^q \gamma_i U_{ij}$$

où $\gamma_1 \geq \gamma_2 \geq \dots \geq \gamma_p$ sont les valeurs propres de Σ_{11} chacune ayant multiplicité q et les U_{ij} sont des variables aléatoires khi-deux non centrées, indépendantes, avec paramètres de non-centralisé δ_{ij}^2 , où les δ_{ij} sont les éléments du vecteur

$(1+k)^{-1/2}P'(\Sigma_{11} \otimes \Sigma_{22})^{-1/2}\text{vec}(A)$ et où P est la matrice formée des vecteurs propres normalisés de $(\Sigma_{11} \otimes \Sigma_{22})^{-1/2}(I_p \otimes \Sigma_{22}^{-1})(\Sigma_{11} \otimes \Sigma_{22})^{-1/2}$

Preuve : La loi asymptotique, lorsque $n \rightarrow \infty$, de nRI est la loi de

$$\frac{n(\text{vec}(S_{21}))'(I_p \otimes \Sigma_{22}^{-1})(\text{vec}(S_{21}))}{\text{tr}\Sigma_{11}}$$

Par le lemme 5.1 de Cléroux et Ducharme (1989), c'est une forme quadratique de variables normales (asymptotiques) avec moyennes non-nulles et il suit que $n(\text{vec}(S_{21}))'(I_p \otimes \Sigma_{22}^{-1})(\text{vec}(S_{21})) \xrightarrow{l} (1+k) \sum_{i=j}^p \sum_{j=1}^q \gamma_i \chi_1^2(\delta_{ij}^2)$ où les δ_{ij} sont bien ceux de l'énoncé du théorème.

5. Tests d'homogénéité de plusieurs coefficients ρI

Considérons $l \geq 2$ populations indépendantes, ayant chacune une loi elliptique. Pour $j = 1, 2, \dots, l$, la j^{e} population a pour matrice covariance $\Sigma^{(j)} = \begin{pmatrix} \Sigma_{11}^{(j)} & \Sigma_{12}^{(j)} \\ \Sigma_{21}^{(j)} & \Sigma_{22}^{(j)} \end{pmatrix}$ où $\Sigma_{11}^{(j)} : p_j \times p_j$ et $\Sigma_{22}^{(j)} : q_j \times q_j$, et pour paramètre d'aplatissement k_j . Pour la j^{e} population on a $\rho I^{(j)}$. L'objectif de cette section est de tester l'hypothèse

$$H_0 : \rho I^{(1)} = \rho I^{(2)} = \dots = \rho I^{(l)} \quad (= \rho I \text{ non spécifié dans } (0, 1)).$$

A partir de la j^{e} population, $j = 1, 2, \dots, l$, on choisit un échantillon aléatoire de taille n_j et on calcule la matrice empirique de covariance $S^{(j)} = \begin{pmatrix} S_{11}^{(j)} & S_{12}^{(j)} \\ S_{21}^{(j)} & S_{22}^{(j)} \end{pmatrix}$ où $S_{11}^{(j)} : p_j \times p_j$ et $S_{22}^{(j)} : q_j \times q_j$. $RI^{(j)}$ est également calculé. Pour tester H_0 on introduit la statistique

$$A = \sum_{j=1}^l \frac{n_j(RI^{(j)} - \hat{\rho}I)^2}{(\hat{\tau}^{(j)})^2} \quad \text{avec} \quad \hat{\rho}I = \frac{\sum_{j=1}^l \frac{n_j RI^{(j)}}{(\hat{\tau}^{(j)})^2}}{\sum_{j=1}^l \frac{n_j}{(\hat{\tau}^{(j)})^2}} \quad (8)$$

où $(\hat{\tau}^{(j)})^2 = 2(RI^{(j)})^2(1 + \hat{k}_j)(\hat{\sigma}^{(j)})^2$, $(\hat{\sigma}^{(j)})^2$ étant l'estimé de $(\sigma^{(j)})^2$ (donné par (6) avec les indices appropriés pour indiquer la j^{e} population) obtenu en remplaçant $\Sigma^{(j)}$ par $S^{(j)}$ et $\hat{k}_j$ étant un estimateur convergent de k_j (voir Muirhead et Waternaux (1980)). L'approche que nous utilisons ici est semblable à celle de

Rao (1965), p. 364 pour tester l'homogénéité d'un ensemble de coefficients de corrélation simple.

Soient $n = \sum_{j=1}^l n_j$ et $\Pi_j = \frac{n_j}{n}$. Nous démontrons le théorème suivant :

Théorème 4. – Sous H_0 , si $n \rightarrow \infty$ et si $\Pi_1, \Pi_2, \dots, \Pi_l$ deviennent constant, alors

$$A = \sum_{j=1}^l \frac{n_j(RI^{(j)} - \hat{\rho}I)^2}{(\hat{\tau}^{(j)})^2} = n \sum_{j=1}^l \frac{\Pi_j(RI^{(j)} - \hat{\rho}I)^2}{(\hat{\tau}^{(j)})^2} \xrightarrow{l} \chi_{l-1}^2$$

Preuve : Sous H_0 on peut écrire

$$A = n \sum_{j=1}^l \frac{\Pi_j(RI^{(j)} - \hat{\rho}I)^2}{(\hat{\tau}^{(j)})^2} = n \sum_{j=1}^l \frac{(RI^{(j)} - \rho I)^2}{(\hat{\tau}^{(j)})^2} - n(\hat{\rho}I - \rho I)^2 \sum_{j=1}^l \frac{\Pi_j}{(\hat{\tau}^{(j)})^2}.$$

Puisque $\tau^{(j)}$ est continu, nous avons

$$n \sum_{j=1}^l \frac{\Pi_j(RI^{(j)} - \rho I)^2}{(\hat{\tau}^{(j)})^2} - n \sum_{j=1}^l \frac{\Pi_j(RI^{(j)} - \rho I)^2}{(\tau^{(j)})^2} \xrightarrow{p} 0$$

et

$$n(\hat{\rho}I - \rho I)^2 \sum_{j=1}^l \frac{\Pi_j}{(\hat{\tau}^{(j)})^2} - n(\rho I^* - \rho I)^2 \sum_{j=1}^l \frac{\Pi_j}{(\tau^{(j)})^2} \xrightarrow{p} 0$$

où ρI^* est $\hat{\rho}I$ avec $\tau^{(j)}$ à la place de $\hat{\tau}^{(j)}$. Donc la loi asymptotique de A est la même que celle de

$$\begin{aligned} A^* &= n \sum_{j=1}^l \frac{\Pi_j(RI^{(j)} - \rho I)^2}{(\tau^{(j)})^2} - n(\rho I^* - \rho I)^2 \sum_{j=1}^l \frac{\Pi_j}{(\tau^{(j)})^2} \\ &= n \sum_{j=1}^l \frac{(RI^{(j)} - \rho I)^2}{(c^{(j)})^2} - n(\rho I^* - \rho I)^2 \sum_{j=1}^l \frac{1}{(c^{(j)})^2} \end{aligned}$$

$$\text{où } (c^{(j)})^2 = \frac{(\tau^{(j)})^2}{\Pi_j} \text{ et } \rho I^* = \frac{\sum_{j=1}^l \frac{RI^{(j)}}{(c^{(j)})^2}}{\sum_{j=1}^l \frac{1}{(c^{(j)})^2}}.$$

Soit $Y^{(j)} = n^{\frac{1}{2}}(RI^{(j)} - \rho I)/c^{(j)}$, $j = 1, 2, \dots, l$. Par le théorème 1 et sous H_0 , lorsque $n \rightarrow \infty$, $Y^{(j)} \xrightarrow{l} N(0, 1)$ et

$$A^* = \sum_{j=1}^l (Y^{(j)})^2 - \frac{1}{\sum_{j=1}^l \frac{1}{(c^{(j)})^2}} \left[\sum_{j=1}^l \frac{Y^{(j)}}{c^{(j)}} \right]^2$$

est une forme quadratique qui s'écrit $Q = Y'FY$ où $Y = (Y^{(1)}, Y^{(2)}, \dots, Y^{(l)})'$ est $N(0, I_l)$ et où

$$F = \begin{bmatrix} 1 - \frac{h}{(c^{(1)})^2} & \frac{-h}{c^{(1)}c^{(2)}} & \cdots & \frac{-h}{c^{(1)}c^{(l)}} \\ \frac{-h}{c^{(2)}c^{(1)}} & 1 - \frac{h}{(c^{(2)})^2} & \cdots & \frac{-h}{c^{(2)}c^{(l)}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{-h}{c^{(l)}c^{(1)}} & \frac{-h}{c^{(l)}c^{(2)}} & \cdots & 1 - \frac{h}{(c^{(l)})^2} \end{bmatrix}$$

avec $h = \left[\sum_{j=1}^l \frac{1}{(c^{(j)})^2} \right]^{-1}$. On voit facilement que F est idempotente avec $\text{rang}(F) = \text{tr}(F) = l - 1$. En conséquence A^* (et A) est asymptotiquement χ_{l-1}^2 . Cela complète la preuve du théorème.

En pratique on calcule A en utilisant la formule

$$A = \sum_{j=1}^l \frac{n_j (RI^{(j)})^2}{(\hat{\tau}^{(j)})^2} - \frac{1}{\sum_{j=1}^l \frac{n_j}{(\hat{\tau}^{(j)})^2}} \left[\sum_{j=1}^l \frac{n_j RI^{(j)}}{(\hat{\tau}^{(j)})^2} \right]^2$$

et l'hypothèse H_0 est rejetée au niveau α si et seulement si $A > \chi_{l-1; \alpha}^2$, le $100(1 - \alpha)^e$ centile de χ_{l-1}^2 . Si H_0 est acceptée, l'estimateur naturel de ρI est $\hat{\rho} I$ donné par (8). Il est asymptotiquement sans biais et sa variance asymptotique est

$$\sigma^2(\hat{\rho} I) = \left[\sum_{j=1}^l \frac{n_j}{(\hat{\tau}^{(j)})^2} \right]^{-1}$$

Exemple 5.1

Considérons les données sur les Iris telles qu'on les trouve dans Kendall, Stuart et Ord (1983, p.375). Sur 50 Iris Versicolor (groupe 1), 50 Iris Virginica (groupe 2) et 50 Iris Setosa (groupe 3), on a mesuré : la longueur du sépale, la largeur du sépale, la longueur du pétale et la largeur du pétale. Les données sont

exprimées en cm. Soient $X^{(1)}$ le vecteur 2×1 des dimensions du sépale et $X^{(2)}$ le vecteur 2×1 des dimensions du pétale et supposons que l'on s'intéresse à la structure de redondance entre $X^{(1)}$ et $X^{(2)}$ pour chaque groupe. Plus précisément, supposons que l'on veuille tester l'hypothèse $H_0 : \rho I^{(1)} = \rho I^{(2)} = \rho I^{(3)} (= \rho I \text{ non-spécifié dans } (0, 1))$.

On a $p_j = q_j = 2$ pour $j = 1, 2, 3$; $n_j = 50$ pour $j = 1, 2, 3$ et $n = 150$. Les paramètres d'aplatissement sont estimés par $\hat{k}_1 = -0.138$, $\hat{k}_2 = -0.041$, $\hat{k}_3 = 0.220$ et les autres résultats numériques sont donnés dans la Table 5.1. Puisque $A > \chi_{2;0.01}^2 = 9.21$, H_0 est rejeté au niveau 0.01.

Pour tester l'hypothèse $H'_0 : \rho I^{(1)} = \rho I^{(2)}$, ($= \rho I$ non-spécifié dans $(0, 1)$) on calcule les résultats de la Table 5.2. Puisque $A < \chi_{1;0.05}^2 = 3.84$, H'_0 est acceptée au niveau 0.05.

Table 5.1
Résultats numériques pour tester $H_0 : \rho I^{(1)} = \rho I^{(2)} = \rho I^{(3)} (= \rho I)$.

$RI^{(1)}$	$RI^{(2)}$	$RI^{(3)}$	A	$\hat{\rho I}$	$\hat{\sigma}^2(\hat{\rho I})$
0.540	0.665	0.087	35.058	0.446	0.0017

Table 5.2
Résultats numériques pour tester $H'_0 : \rho I^{(1)} = \rho I^{(2)} (= \rho I)$.

A	$\hat{\rho I}$	$\hat{\sigma}^2(\hat{\rho I})$
1.632	0.613	0.0023

Exemple 5.2

Les données de cet exemple proviennent de BMDP Statistical Software (1981, p. 38). On a 188 vecteurs de taille 8 et les variables sont l'âge, la taille, le poids, la prise de la pilule anticonceptionnelle, le taux de cholestérol, d'albumine, de calcium et d'acide urique. 7 vecteurs sont éliminés à cause de données manquantes. Soit $X^{(1)}$ le vecteur 3×1 formé des trois premières variables (physiques) et $X^{(2)}$ le vecteur 4×1 formé des quatre dernières (chimie sanguine) et supposons que nous nous intéressons à la structure de redondance entre $X^{(1)}$ et $X^{(2)}$ pour deux groupes de patientes : les 89 patientes ne prenant pas la pilule (birthpill = 1) et les 92 patientes prenant la pilule (birthpill = 2).

Plus précisément on veut tester $H_0 : \rho I^{(1)} = \rho I^{(2)} (= \rho I \text{ non spécifié dans } (0, 1))$. Nous avons $p_j = 3$, $q_j = 4$ pour $j = 1, 2$, $n_1 = 89$, $n_2 = 92$ et $n = 181$. Les paramètres d'aplatissement sont estimés par $\hat{k}_1 = -0.042$ et $\hat{k}_2 = 0.284$ et les résultats numériques sont donnés dans la Table 5.3.

Table 5.3
Résultats numériques pour tester $H_0 : \rho I^{(1)} = \rho I^{(2)} (= \rho I)$.

$RI^{(1)}$	$RI^{(2)}$	A	$\hat{\rho}I$	$\hat{\sigma}^2(\hat{\rho}I)$
0.157	0.172	0.030	0.163	0.0019

Puisque $A < \chi_{1;0.05}^2 = 3.84$, H_0 est acceptée au niveau 0.05.

6. Test de l'hypothèse que plusieurs ρI sont nuls

Le contexte de cette section est le même qu'à la section précédente sauf l'hypothèse à tester qui est $H_0 : \rho I^{(1)} = \rho I^{(2)} = \dots = \rho I^{(l)} = 0$.

Dans la situation univariée, pour tester l'hypothèse que plusieurs coefficients de corrélation simple sont nuls, $H_0 : \rho_1 = \rho_2 = \dots = \rho_l = 0$, Donner et Rosner (1980) ont proposé une procédure basée sur la statistique

$$\hat{\rho} = \frac{\sum_{j=1}^l (n_j - 1) r_j}{\sum_{j=1}^l (n_j - 1)} \quad (9)$$

où r_j est le coefficient empirique de corrélation simple du j^e groupe et n_j est la taille de l'échantillon. Ils comparent $t_s = \frac{(N - l - 1)^{\frac{1}{2}} \hat{\rho}}{(1 - \hat{\rho}^2)^{\frac{1}{2}}}$ au point critique de la

loi t avec $N - l - 1$ degrés de libertés, où $N = \sum_{j=1}^l (n_j - 1)$, et H_0 est rejetée au niveau α si et seulement si $|t_s| > t_{N-l-1;\alpha}$. Leur test est validé par des simulations de Monte Carlo.

Par analogie à (9) nous proposons une procédure basée sur la statistique $\hat{\rho}I^0 = \frac{1}{n} \sum_{j=1}^l n_j RI^{(j)}$ pour tester H_0 , où $n = \sum_{j=1}^l n_j$.

Théorème 4. – Sous H_0 et lorsque $n \rightarrow \infty$ (chaque $n_j \rightarrow \infty$, $j = 1, 2, \dots, l$), la loi asymptotique de $n\rho\hat{I}^0$ est la loi $\sum_{j=1}^l \sum_{s=1}^{p_j q_j} b_{js} W_{js}^2$ où les W_{js} sont iid $N(0, 1)$,

$b_{js} = \frac{\varepsilon_s^{(j)}(1 + k_j)}{\text{tr}(\Sigma_{11}^{(j)})}$ et les $\varepsilon_s^{(j)}$ sont les valeurs propres de $\Sigma_{11}^{(j)} \otimes I_{q_j}$.

Preuve : A partir des lois asymptotiques de chaque $n_j RI^{(j)}$ lorsque $\rho I^{(j)} = 0$, la loi asymptotique de $n\hat{\rho}I^0$ est la loi de

$$\begin{aligned} T &= \sum_{j=1}^l \frac{1+k_j}{\text{tr}(\Sigma_{11}^{(j)})} \sum_{s=1}^{p_j} \sum_{r=1}^{q_j} \delta_s^{(j)} (U_{sr}^{(j)})^2 \\ &= \sum_{j=1}^l \frac{1+k_j}{\text{tr}(\Sigma_{11}^{(j)})} \sum_{s=1}^{p_j q_j} \varepsilon_s^{(j)} W_{js}^2 \\ &= \sum_{j=1}^l \sum_{s=1}^{p_j q_j} b_{js} W_{js}^2 \end{aligned}$$

où les W_{js} sont iid $N(0, 1)$ puisque les coefficients $RI^{(j)}$ sont indépendants et où $b_{js} = \frac{\varepsilon_s^{(j)}(1+k_j)}{\text{tr}(\Sigma_{11}^{(j)})}$. Pour chaque $j = 1, 2, \dots, l$, puisque $\delta_1^{(j)}, \delta_2^{(j)}, \dots, \delta_{p_j}^{(j)}$ sont les valeurs propres de $\Sigma_{11}^{(j)}$, il suit que $\varepsilon_1^{(j)}, \varepsilon_2^{(j)}, \dots, \varepsilon_{p_j q_j}^{(j)}$ sont les valeurs propres $\Sigma_{11}^{(j)} \otimes I_{q_j}$, chacune ayant une multiplicité q_j . Cela complète la preuve du théorème.

L'hypothèse $H_0 : \rho I^{(1)} = \rho I^{(2)} = \dots = \rho I^{(l)} = 0$ est donc rejetée au niveau α si et seulement si $n\hat{\rho}I^0 > T_\alpha$ où T_α est le $100(1 - \alpha)^e$ centile de la loi de T . En pratique, pour calculer T_α , on emploie l'algorithme de Imhof. Puisque les paramètres sont habituellement inconnus on les remplace par les estimateurs usuels : $\Sigma_{11}^{(j)}$ et $\Sigma_{22}^{(j)}$ sont remplacés par $S_{11}^{(j)}$ et $S_{22}^{(j)}$, $\varepsilon_1^{(j)}, \varepsilon_2^{(j)}, \dots, \varepsilon_{p_j q_j}^{(j)}$ sont remplacés par $\hat{\varepsilon}_1^{(j)}, \hat{\varepsilon}_2^{(j)}, \dots, \hat{\varepsilon}_{p_j q_j}^{(j)}$, les valeurs propres de $S_{11}^{(j)} \otimes I_{q_j}$, et k_j remplacé par un estimateur convergent $\hat{k}_j$ (voir Muirhead et Waternaux (1980)).

Exemple 6.1

Considérons encore une fois les données sur les Iris Setosa (groupe 3 de l'Exemple 5.1) et soit à tester l'hypothèse $H_0 : \rho I^3 = 0$. Nous avons $p_3 = q_3 = 2$, $n_3 = 50$, $RI^{(3)} = 0.087$, $n_3 RI^{(3)} = 4.350$. En employant l'algorithme de Imhof, on obtient que le point critique à 5% de la loi asymptotique de $n_3 RI^{(3)}$ est 6.69. Donc H_0 ne peut pas être rejetée.

Exemple 6.2

Revenons aux données de l'Exemple 5.2 et testons l'hypothèse $H_0 : \rho I^{(1)} = \rho I^{(2)} = 0$. Nous avons $p_j = 3$, $q_j = 4$ pour $j = 1, 2$, $n_1 = 89$, $n_2 = 92$, $n = 181$ et calculons $RI^{(1)} = 0.157$, $RI^{(2)} = 0.172$ et $n\hat{\rho}I^0 = 29.797$. Le point critique à 5% de la loi asymptotique de $n\hat{\rho}I^0$ est 16.04. On rejette donc H_0 au niveau 0.05.

7. Conclusion

Dans cet article nous nous sommes intéressés à des hypothèses d'homogénéité entre plusieurs indices de redondance dans un contexte de lois elliptiques qui inclut

la loi normale multivariée, la loi normale multivariée contaminée et plusieurs autres lois. Des études de robustesse sont donc possibles sans trop de difficultés.

D'autre part, à partir de (3) et (5), il serait facile de définir des indices de redondance partielle en remplaçant Σ et S par des matrices de covariance partielle. Les résultats des Sections 2 à 6 s'étendraient directement à ces nouveaux indices.

Il est finalement possible d'obtenir la loi conjointe de plusieurs indices de redondance quand le vecteur X est partitionné en plus de deux sous-vecteurs. Cela permet de tester, par exemple, l'hypothèse d'égalité de deux indices de redondance quand les populations ne sont pas indépendantes. Le lecteur intéressé est invité à lire la thèse de doctorat de Lazraq (1989).

Remerciements

Les auteurs expriment leur reconnaissance au comité de rédaction pour leur précieux commentaires sur une précédente version de cet article.

Annexe

On rappelle brièvement la définition d'une loi elliptique. Pour plus de détails, voir par exemple Muirhead (1982), sect. 1.5. Le vecteur aléatoire $X : m \times 1$ possède une loi elliptique avec paramètres $\mu : m \times 1$ et $V : m \times m$ si sa fonction de densité est de la forme $C_m(\det V)^{-1/2}h((x - \mu)'V^{-1}(x - \mu))$ où C_m est une constante de normalisation, $h(\cdot)$ est une fonction qui ne dépend pas de m et V est définie positive. On écrira $X : E_m(\mu, V)$ et on dira que la loi de X est dans la classe des lois elliptiques.

Si $X : E_m(\mu, V)$, on peut montrer que X a pour vecteur moyenne μ et pour matrice de covariance αV pour une certaine constante α . La fonction caractéristique de X est donnée par $Q(t) = e^{it'\mu}\Psi(t'Vt)$ pour une certaine fonction $\Psi(\cdot)$. De plus, $\alpha = -2\Psi'(0)$.

Si $X : E_m(\mu, V)$, on peut montrer que les lois marginales des m composantes de X possèdent toutes un coefficient d'asymétrie qui est nul et un coefficient d'aplatissement donné par $3k$ où $k = \frac{\Psi''(0)}{[\Psi'(0)]^2} - 1$. Le coefficient k est appelé paramètre d'aplatissement. Pour une loi normale multivariée, $k = 0$.

Bibliographie

- Billingsley P. (1968) *Convergence of Probability Measures*, John Wiley and Sons, New York.
- BMDP Statistical Software (1981) W.J. Dixon, Chief Editor, University of California Press, Berkeley.
- Cléroux R. and Ducharme G.R. (1989) Vector Correlation for Elliptical Distributions, *Communications in Statistics, Theory and Methods*, 18, 1441-1454.

- Coxhead P. (1974) Measuring the Relationship Between two Sets of Variables, *British Journal of Mathematical and Statistical Psychology*, 27, 205-212.
- Cramer E.M. (1974) A Generalization of Vector Correlation and its Relation to Canonical Correlation, *Multivariate Behavioral Research*, 9, 347-352.
- Cramer E.M. and Nicewander W.A. (1979) Some Symmetric, Invariant Measures of Multivariate Association, *Psychometrika*, 49, 403-423.
- Donner A. and Rosner B. (1980) On Inferences Concerning a Common Correlation Coefficient, *Applied Statistics*, 29, 69-76.
- Dwyer P.S. (1967) Some Applications of Matrix Derivatives in Multivariate Analysis, *Journal of the American Statistical Association*, 62, 607-625.
- Escoufier Y. (1973) Le traitement des variables vectorielles, *Biometrics*, 29, 751-760.
- Gleason T.C. (1976) On Redundancy in Canonical Analysis, *Psychological Bulletin*, 83, 1004-1006.
- Hotelling H. (1936) Relations Between Two Sets of Variables, *Biometrika*, 28, 321-377.
- Imhof P. (1961) Computing the Distribution of Quadratic Forms in Normal Variates, *Biometrika*, 48, 419-426.
- Kendall M., Stuart A. and Ord J.K. *The Advanced Theory of Statistics*, vol. 3, (MacMillan Publishing, New York, 1983, 4th. ed.).
- Koerts J and Abrahamse A.P.J. (1969) On the Theory and Application of the General linear Model, *Rotterdam University Press, Rotterdam*.
- Kshirsagar A.M. (1969) Correlation Between Two Vector Variables, *Journal of the Royal Statistical Society, Series B*, 31, 477-485.
- Lazraq A. (1989) Inférences sur plusieurs mesures de liaison entre deux vecteurs aléatoires et algorithmes de sélection ou de variables. Thèse de doctorat. Université de Montréal.
- Lazraq A. and Cléroux R. (1988) Un algorithme pas à pas de sélection de variables en régression linéaire multivariée, *Statistique et Analyse des données*, 13, 39-58.
- Lingoes J.C. and Schonemann P.H. (1974) Alternative Measures of Fit for the Schonemann-Carroll Matrix-Fitting Algorithm, *Psychometrika*, 39, 423-427.
- Masuyama M. (1939) Correlation Between Tensor Quantities, *Proceedings of the Physico-Mathematical Society of Japan, Series 3*, 31, 638-647.
- Masuyama M. (1941) Correlation Between Two Sets of Complex Vectors, *Proceedings of the Physico-Mathematical Society of Japan, Series 3*, 33, 918-924.
- Muirhead R.J. (1982) *Aspects of Multivariate Statistical Theory*, John Wiley and Sons, New York.
- Muirhead R.J. and Waternaux C.M. (1980) Asymptotic distributions in Canonical Correlation Analysis and Other Multivariate Procedures for Nonnormal Populations, *Biometrika*, 67, 31-43.

- Press S.J. (1972) *Applied Multivariate Analysis*, Holt, Rinehart and Winston, Chicago.
- Ramsay J.O., Ten Berge J. and Styan G. (1984) Matrix Correlation, *Psychometrika*, 49, 403-423.
- Rao C.R. (1965) *Linear Statistical Inference and its Applications*, John Wiley and Sons, New York.
- Robert P. and Escoufier Y. (1976) A Unifying Tool for Linear Multivariate Statistical Methods : the RV-Coefficient, *Applied Statistics*, 25, 257-265.
- Roseboom W.W. (1965) Linear Correlation Between Sets of Variables, *Psychometrika*, 30, 57-71.
- Seber G.A.F. (1984) *Multivariate Observations*, John Wiley and Sons, New York.
- Shaffer J.P. and Gillo M.W. (1974) A multivariate Extension of the Correlation Ratio, *Educational and Psychological Measurements*, 34, 521-524.
- Sibson R. (1978) Studies in the Robustness of Multidimensional Scaling Procrustes Statistics, *Journal of the Royal Statistical Society, Series B*, 40, 234-235.
- Stephens M.A. (1979) Vector Correlation, *Biometrika*, 66, 41-48.
- Stewart D and Love W (1968) A General Canonical Correlation Index, *Psychological Bulletin*, 70, 160-163.