Statistique et analyse des données

Aziz Lazraq
Robert Cléroux

Étude comparative de différentes mesures de liaison entre deux vecteurs aléatoires et tests d'indépendance

Statistique et analyse des données, tome 13, n° 1 (1988), p. 15-38
<http://www.numdam.org/item?id=SAD_1988__13_1_15_0>

Statistique et Analyse des Données
1988 - Vol. 13 n° 1 - p.15-38

ETUDE COMPARATIVE DE DIFFERENTES MESURES DE LIAISON ENTRE DEUX VECTEURS ALEATOIRES ET TESTS D'INDEPENDANCE

Aziz Lazraq et Robert Cléroux

Département d'informatique et recherche opérationnelle, Université de Montréal
C.P. 6128, Succ. A, Montréal, Québec, Canada, H3C 3J7

Résumé : *Nous nous intéressons à différentes mesures de liaison entre deux ensembles de variables. Complétant les travaux de Cramer et Nicewander [6] nous effectuons une étude comparative de différentes telles mesures et les ordonnons partiellement. Nous étudions également leurs propriétés. Finalement, deux d'entre elles sont utilisées pour tester l'hypothèse d'indépendance entre deux vecteurs aléatoires.*

Abstract : *In this paper different measures of redundancy or association between sets of variables are considered. Following and completing previous work by Cramer and Nicewander [6], several such measures are compared and partially ordered. Their properties are also studied. Finally two of them are used to test the hypothesis of independence between two random vectors.*

Mots clés : *Mesures de liaison, redondance, association, corrélation vectorielle. Tests d'indépendance.*

Indices de classification STMA : *06-010; 06-040; 06-900.*

Manuscrit reçu le 18.5.87, révisé le 2.5.88

1. INTRODUCTION

Dans cet article nous nous intéressons à différentes mesures de corrélation, de redondance ou d'association entre deux vecteurs aléatoires, deux ensembles de variables ou deux tableaux d'observations. Depuis les travaux de Hotelling [8], plusieurs telles mesures furent proposées dans différents contextes. Elles furent regroupées en deux ensembles par Cramer et Nicewander [6]: les mesures de redondance qui sont associées à la prédiction d'un ensemble de variables par un autre et les mesures d'association qui sont des généralisations du concept de coefficient de corrélation à deux ensembles de variables. Dans la suite du texte nous utiliserons "mesures de liaison" pour signifier à la fois les mesures de redondance et les mesures d'association.

Outre [6] et [8], plusieurs auteurs se sont intéressés à des mesures de liaison entre vecteurs aléatoires: Roseboom [19], Stewart et Love [22], Kshirsagar [11], Coxhead [4], Cramer [5], Shaffer et Gillo [20] ainsi que Ramsay, ten Berge et Styan [16]. La plupart de ces mesures sont fonctions des corrélations canoniques entre les deux vecteurs. Dans un contexte différent, Masuyama [13] et [14] a également étudié la liaison entre vecteurs. Finalement, Escoufier [7] et Stephens [21] ont proposé des mesures de liaison possédant la propriété que deux ensembles de vecteurs seraient parfaitement corrélés si une transformation orthogonale faisait coïncider l'un avec l'autre.

Dans cet article, en nous inspirant de Cramer et Nicewander [6], nous effectuons une étude comparative de différentes mesures de liaison (2 mesures de redondance et 7 mesures d'association), nous les ordonnons partiellement, puis nous en choisissons une dans chaque groupe pour tester l'indépendance entre deux vecteurs aléatoires. Dans la section 2 nous rappelons les contextes de la régression linéaire multivariée ainsi que de l'analyse canonique et nous posons la notation. Les différentes mesures de liaison sont introduites dans la section 3 et ordonnées dans la section 4. Les tests d'indépendance entre vecteurs aléatoires sont proposés dans la section 5.

2. RAPPELS ET NOTATIONS

Lorsque l'on cherche à expliquer une variable aléatoire par une combinaison linéaire des composantes d'un vecteur explicatif, la qualité de la régression se mesure entre autre par le coefficient de corrélation multiple. Dans le cas d'une régression multivariée, la qualité de l'ajustement sera définie à partir d'une mesure de liaison entre deux vecteurs aléatoires. Nous effectuons un bref rappel des modèles de régression multivariée et de corrélation canonique afin d'établir le contexte de travail ainsi que la notation utile.

2.1. La régression multivariée

Soit un vecteur aléatoire $\underline{Y}$: px1 que l'on cherche à expliquer linéairement à partir d'un vecteur explicatif $\underline{X}$: qx1. Le modèle de la régression linéaire multivariée s'écrit

$$Y = X B + E \tag{2.1}$$

où

(i) Y est un tableau nxp de n mesures prises sur le vecteur $\underline{Y}$, représentant les observations sur n individus et p variables,

(ii) X est un tableau nxq représentant les observations sur q variables et sur les mêmes individus,

(iii) B est la matrice qxp des coefficients de régression,

(iv) E est une matrice nxp des mesures des résidus (de ce qui reste quand on explique Y par XB).

On suppose que les matrices Y et X sont centrées et de plein rang et que $p \leq q$ (ce n'est pas une restriction). Lorsque l'on voudra faire de l'inférence on devra supposer également que les lignes de E sont indépendantes, identiquement distribuées et p-normales.

L'estimation de la matrice B des coefficients de régression est obtenue en minimisant la somme des carrés des erreurs $tr(E'E)$:

$$\frac{d}{dB}\, tr(E'E) = 0 \Longrightarrow \hat{B} = (X'X)^{-1}X'Y.$$

Si on pose $S_{xx} = X'X$, $S_{xy} = X'Y$ et $S_{yy} = Y'Y$ on peut écrire

$$\hat{B} = S_{xx}^{-1} S_{xy} . \tag{2.2}$$

Si on pose également $S_{\hat{y}\hat{y}} = \hat{Y}'\hat{Y} = (X\hat{B})'(X\hat{B})$ on obtient

$$S_{\hat{y}\hat{y}} = \hat{B}' X'X \hat{B} = Y'X(X'X)^{-1} X'Y = S_{yx} S_{xx}^{-1} S_{xy} . \tag{2.3}$$

Finalement, notons que

a) pour chaque i, $1 \leq i \leq p$, la i^e colonne de $\hat{B}$ peut également être obtenue par une régression linéaire multiple de la i^e composante de $\underline{Y}$ sur le vecteur $\underline{X}$,

b) la diagonale de la matrice S_{yy} contient les variances (à une constante multiplicative près) des composantes du vecteur $\underline{Y}$. Le i^e élément de la diagonale est noté $SS_{y(i)}$, i=1,2,...,p,

c) de même, la diagonale de la matrice $S_{\hat{y}\hat{y}}$ contient les variances (à une constante multiplicative près) des composantes du vecteur expliqué $\underline{\hat{Y}}$. Le i^e élément de cette diagonale est noté $SS_{\hat{y}(i)}$, i=1,2,...,p,

d) pour chaque i, $1 \leq i \leq p$, le carré du coefficient de corrélation multiple est

$$R^2_{y(i)\, x_1, x_2, \ldots, x_q} = \frac{SS_{\hat{y}(i)}}{SS_{y(i)}} \tag{2.4}$$

le rapport entre la variance expliquée et la variance totale.

2.2. Corrélations canoniques

Supposons que le problème consiste à trouver une combinaison linéaire $v'\underline{\hat{Y}}$ qui explique au mieux la combinaison linéaire $v'\underline{Y}$, c'est-à-dire qui maximise le rapport entre la variance expliquée $var\,(v'\underline{\hat{Y}}) = v'S_{\hat{y}\hat{y}}v$ et la variance totale $var\,(v'\underline{Y}) = v'S_{yy}v$. On cherche donc

$$\max_{v} \frac{v'S_{\hat{y}\hat{y}}v}{v'S_{yy}v} = \max_{v} \frac{v'S_{yx}S_{xx}^{-1}S_{xy}v}{v'S_{yy}v} . \tag{2.5}$$

Ce problème peut être résolu par la méthode des multiplicateurs de Lagrange en posant $v'S_{yy}v = 1$ et en cherchant à dériver

$$L(\lambda, v) = v'S_{yx} S_{xx}^{-1} S_{xy} v - \lambda(v'S_{yy} v - 1) .$$

On réalise facilement que c'est un problème de valeurs et de vecteurs propres avec

$$S_{\hat{y}\hat{y}} v_i = \hat{\rho}_i^2 S_{yy} v_i , \; i=1,2,..., p . \tag{2.6}$$

Les quantités $\hat{\rho}_i^2$ sont les valeurs propres de $S_{yy}^{-1} S_{\hat{y}\hat{y}}$ et les vecteurs v_i sont les vecteurs propres correspondants, normalisés de telle façon que $v_i' S_{yy} v_i = 1$, i=1,2,..., p. Les racines carrés positives (notées $\hat{\rho}_1, \hat{\rho}_2, \ldots, \hat{\rho}_p$) de $\hat{\rho}_1^2, \hat{\rho}_2^2, \ldots, \hat{\rho}_p^2$ sont appelées corrélations canoniques et sont arrangées en ordre décroissant: $\hat{\rho}_1 \geq \hat{\rho}_2 \geq \cdots \geq \hat{\rho}_p$. Il est connu que $\hat{\rho}_1 \leq 1$ et $\hat{\rho}_p \geq 0$. Les vecteurs $v_1, v_2, \ldots, v_p$ sont appelés vecteurs canoniques.

L'équation (2.6) peut également s'écrire sous la forme matricielle

$$S_{yx} S_{xx}^{-1} S_{xy} V = S_{yy} V \Lambda \tag{2.7}$$

où V est la matrice dont la i^e colonne est le vecteur canonique v_i, i=1,2..., p et où Λ = diag $(\hat{\rho}_1^2, \hat{\rho}_2^2, \ldots, \hat{\rho}_p^2)$. Si l'on prémultiplie (2.7) par V' on obtient

$$V' S_{yx} S_{xx}^{-1} S_{xy} V = V' S_{yy} V \Lambda = \Lambda . \tag{2.8}$$

En outre, si l'on pose $G = V^{-1}$, on peut écrire

$$S_{\hat{y}\hat{y}} = S_{yx} S_{xx}^{-1} S_{xy} = G' \Lambda G \tag{2.9}$$

et

$$S_{yy} = G'G . \tag{2.10}$$

Ces résultats seront utiles par la suite.

3. MESURES DE LIAISON ENTRE VECTEURS

Dans cette section nous étudions 9 mesures de liaison entre les vecteurs $\underline{Y}$ et $\underline{X}$. Nous verrons, au paragraphe (3.10) lesquelles sont des mesures de redondance et lesquelles sont des mesures d'association.

3.1. La mesure de Hotelling [8] - Cramer [5]

Cette mesure, qui est une généralisation de (2.4) est définie par

$$RV_1 = \frac{|S_{\hat{y}\hat{y}}|}{|S_{yy}|} = \frac{|S_{yy}^{-1} S_{yx} S_{xx}^{-1} S_{xy}|}{|S_{yy}^{-1} S_{yy}|} = |S_{yy}^{-1} S_{yx} S_{xx}^{-1} S_{xy}| \tag{3.1}$$

où, pour toute matrice carrée C, $|C|$ dénote son déterminant. RV_1 peut également s'écrire sous

la forme

$$RV_1 = \prod_{i=1}^{P} \hat{\rho}_i^2 \,. \tag{3.2}$$

Puisque $|S_{\hat{y}\hat{y}}|$ et $|S_{yy}|$ sont des variances généralisées (Anderson [1]), RV_1 est un rapport de la variance expliquée $|S_{\hat{y}\hat{y}}|$ sur la variance totale $|S_{yy}|$. C'est cependant une mesure pessimiste car

$$RV_1 \leq \min_i \hat{\rho}_i^2 \,. \tag{3.3}$$

Son utilisation n'est donc pas recommandée dans un algorithme de sélection de variables puisque RV_1 nous amène facilement à accepter une liaison peu significative.

Finalement, puisque RV_1 est une fonction des corrélations canoniques, elle est invariante sous les transformations linéaires des variables originales. Elle est également symétrique en ce sens que $RV_1(x, y) = RV_1(y, x)$.

3.2. La mesure de Hotelling [8] - Roseboom [19]

La mesure de Hotelling - Roseboom est également une généralisation de (2.4). Si on écrit $SS_{y(i)} = SS_{\hat{y}(i)} + SS_{e(i)}$ alors (2.4) devient

$$R^2_{y(i)\cdot x_1, x_2, \ldots, x_q} = 1 - \frac{SS_{e(i)}}{SS_{y(i)}} \,. \tag{3.4}$$

Dans un contexte multivarié, si on écrit également $S_{yy} = S_{\hat{y}\hat{y}} + S_{ee}$, alors (3.4) se généralise par

$$RV_2 = 1 - \frac{|S_{ee}|}{|S_{yy}|} = \frac{|S_{yy}| - |S_{ee}|}{|S_{yy}|} = 1 - |S_{yy}^{-1} S_{ee}|$$

$$= 1 - |I - S_{yy}^{-1} S_{yx} S_{xx}^{-1} S_{xy}| = 1 - \prod_{i=1}^{P} (1 - \hat{\rho}_i^2) \,. \tag{3.5}$$

RV_2 est une mesure optimiste puisque $\prod_{i=1}^{P} (1-\hat{\rho}_i^2) \leq 1 - \max_i \hat{\rho}_i^2$, ce qui entraîne $RV_2 \geq \max_i \hat{\rho}_i^2$. Tout comme RV_1, son utilisation n'est pas recommandée dans un algorithme de sélection de variables. D'autre part, RV_2 est une mesure symétrique et invariante sous les transformations linéaires.

3.3. Des mesures proposées par Cramer et Nicewander [6]

En notant que le déterminant $|S|$ d'une matrice S: pxp est un volume (un cube dans l'espace de dimension p) et que la longueur de son arête est obtenue en considérant $|S|^{\frac{1}{p}}$, on définit

$$RV_3 = \frac{|S_{\hat{y}\hat{y}}|^{\frac{1}{p}}}{|S_{yy}|^{\frac{1}{p}}} = \left[\prod_{i=1}^{p} \hat{\rho}_i^2\right]^{\frac{1}{p}} \tag{3.6}$$

et

$$RV_4 = 1 - \frac{|S_{ee}|^{\frac{1}{p}}}{|S_{yy}|^{\frac{1}{p}}} = 1 - \left[\prod_{i=1}^{p} (1-\hat{\rho}_i^2)\right]^{\frac{1}{p}} . \tag{3.7}$$

On voit facilement que $RV_3 \leq [\min_i \hat{\rho}_i^2]^{\frac{1}{p}}$ et que $RV_4 \geq 1 - [\min_i (1-\hat{\rho}_i^2)]^{\frac{1}{p}}$ Ces mesures sont symétriques et invariantes sous les transformations linéaires mais peu recommandables dans un algorithme de sélection de variables.

3.4. Mesures de liaison et distances

Dans ce paragraphe nous écrirons le coefficient de corrélation multiple en termes de distances avant de le généraliser en mesure de liaison au cas multivarié.

Soit $\underline{Y}$ une variable aléatoire et soit $y_1, y_2, \ldots, y_n$ un échantillon aléatoire provenant de $\underline{Y}$. Le carré de la distance euclidienne entre les observations i et j est $d_{ij}^2 = (y_i - y_j)^2$, d'où on tire

$$d^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}^2 = 2n(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2) = 2n\, SS_y .$$

De la même façon on écrit $\hat{d}^2 = 2n\, SS_{\hat{y}}$ et (2.4) devient

$$\frac{SS_{\hat{y}}}{SS_y} = \frac{\hat{d}^2}{d^2} . \tag{3.8}$$

Considérons maintenant $\underline{Y}$ comme un vecteur aléatoire pxl et soit $y_1, y_2, \ldots, y_n$ un échantillon aléatoire tiré de $\underline{Y}$. Le carré de la distance euclidienne entre les individus i et j, par rapport à la métrique M: pxp définie positive est $D_{ij}^2 = (y_i - y_j)' M(y_i - y_j)$ d'où

$$D^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} D_{ij}^2 = tr\, M \sum_{i=1}^{n} \sum_{j=1}^{n} (y_i - y_j)(y_i - y_j)' = 2n\, tr(MS_{yy}) .$$

De la même façon, $\hat{D}^2 = 2n\, tr(MS_{\hat{y}\hat{y}})$. Une mesure de liaison (en termes de la métrique M) entre deux vecteurs aléatoires qui généralise (3.8) est donnée par

$$\frac{\hat{D}^2}{D^2} = \frac{tr(MS_{y\hat{y}})}{tr(MS_{yy})} . \tag{3.9}$$

C'est le rapport du cumul des carrés des distances entre les paires d'individus dans l'espace défini par les vecteurs prédits et du cumul des carrés des distances entre les paires d'individus dans l'espace défini par les vecteurs observés.

3.5. La mesure de Stewart et Love [22]

En posant M=I dans (3.9) nous avons

$$RV_S = \frac{tr\, S_{\hat{y}\hat{y}}}{tr\, S_{yy}}$$

qui peut également s'écrire sous la forme

$$RV_S = \frac{\sum_{i=1}^{p} \sigma_{y(i)}^2 R_{y(i) \cdot x_1, x_2, \ldots, x_q}^2}{\sum_{i=1}^{p} \sigma_{y(i)}^2} . \tag{3.10}$$

Cette mesure s'exprime donc comme une moyenne pondérée (par les variances) des carrés des coefficients de corrélation multiple entre les composantes du vecteur à prédire et le vecteur de prédiction. RV_S s'écrit également, par (2.9) et (2.10), comme

$$RV_S = \frac{tr\,(S_{yx} S_{xx}^{-1} S_{xy})}{tr\, S_{yy}} = \frac{tr\,(G' \Lambda G)}{tr\,(G'G)} . \tag{3.11}$$

C'est une mesure que l'on peut avantageusement utiliser dans un algorithme de sélection de variables. Cependant elle n'est pas symétrique en y et x ni invariante dans les

transformations linéaires des variables.

3.6. Autre mesure de Cramer et Nicewander [6]

Si on pose $M = S_{yy}^{-1}$ dans (3.9) il vient

$$RV_6 = \frac{tr(S_{yy}^{-1}\, S_{\hat{y}\hat{y}})}{tr(S_{yy}^{-1}\, S_{yy})} = \frac{1}{p}\, tr(S_{yy}^{-1}\, S_{yx}\, S_{xx}^{-1}\, S_{xy}) = \frac{1}{p} \sum_{i=1}^{P} \hat{\rho}_i^2 , \tag{3.12}$$

la moyenne arithmétique des corrélations canoniques. C'est une mesure symétrique qui tient compte de toutes les corrélations canoniques et qui peut être utilisée avec profit dans un algorithme de sélection de variables. Elle est également invariante.

3.7. La mesure de Coxhead [4] et de Shaffer et Gillo [20]

Cette mesure de liaison, proposée indépendamment par Coxhead et par Shaffer et Gillo prend $M = S_{ee}^{-1}$. La distance D_{ij}^2 devient la distance de Mahalanobis $D_{ij}^2 = (y_i - y_j)'\, S_{ee}^{-1} (y_i - y_j)$ et (3.9) s'écrit

$$RV_7 = \frac{tr(S_{ee}^{-1}\, S_{\hat{y}\hat{y}})}{tr(S_{ee}^{-1}\, S_{yy})} . \tag{3.13}$$

Montrons que RV_7 est également une fonction des corrélations canoniques. A partir de $S_{yy} = S_{\hat{y}\hat{y}} + S_{ee}$ et de (2.7) on a successivement

$$S_{\hat{y}\hat{y}}\, V = (S_{\hat{y}\hat{y}} + S_{ee})\, V\, \Lambda$$

$$S_{\hat{y}\hat{y}}\, V - S_{\hat{y}\hat{y}}\, V\, \Lambda = S_{ee}\, V\, \Lambda$$

$$S_{\hat{y}\hat{y}}\, V\, (I - \Lambda) = S_{ee}\, V\, \Lambda \text{ (en supposant } \hat{\rho}_i^2 \neq 1, \textit{pour tout } i \text{)}$$

$$S_{\hat{y}\hat{y}}\, V = S_{ee}\, V\, \Lambda\, (I - \Lambda)^{-1} \tag{3.14}$$

$$S_{ee}^{-1}\, S_{\hat{y}\hat{y}} = V\, \Lambda\, (I - \Lambda)^{-1} V^{-1}$$

$$\text{d'où } \quad tr(S_{ee}^{-1}\, S_{\hat{y}\hat{y}}) = tr[\Lambda (I - \Lambda)^{-1}] = \sum_{i=1}^{P} \frac{\hat{\rho}_i^2}{1 - \hat{\rho}_i^2} . \tag{3.15}$$

De la même façon, par (2.7) et (3.14) on a

$$S_{yy}\, V = S_{ee}\, V(I-\Lambda)^{-1}$$

qui conduit à

$$tr(S_{ee}^{-1}\, S_{yy}) = tr(I-\Lambda)^{-1} = \sum_{i=1}^{P} \frac{1}{(1-\rho_i^2)} \,. \tag{3.16}$$

Finalement, par (3.15) et (3.16),

$$RV_7 = \frac{\sum_{i=1}^{P} d_i\, \rho_i^2}{\sum_{i=1}^{P} d_i} \quad avec\ d_i = \frac{1}{1-\rho_i^2} \,. \tag{3.17}$$

On peut également écrire RV_7 en fonction de la moyenne harmonique des $1-\rho_i^2$:

$$RV_7 = 1 - \frac{P}{\sum_{i=1}^{P} \frac{1}{1-\rho_i^2}} \,. \tag{3.18}$$

Notons que $RV_7 \to 1$ quand $\rho_i^2 \to 1$ pour un i, et que cette mesure est symétrique et invariante.

3.8. La mesure d'Escoufier [7]

Le contexte dans lequel cette mesure fut définie est différent du contexte précédent. On se base ici sur la notion de distance entre deux matrices de données. Ainsi si l'on pose $Y = (y_1, y_2, \ldots, y_n)$: pxn et $X = (x_1, x_2, \ldots, x_n)$: qxn, en notant que pour toute matrice carrée E la fonction $f(E) = \|E\| = \sqrt{tr(E'E)}$ est une norme, alors la distance induite entre Y et X est

$$dist(Y, X) = \left| \frac{Y'Y}{\sqrt{tr(Y'Y)^2}} - \frac{X'X}{\sqrt{tr(X'X)^2}} \right| \,.$$

On montre facilement que $dist(Y, X) = \sqrt{2}\,\sqrt{1-RV_8}$ où

$$RV_8 = \frac{tr(S_{yx}\, S_{xy})}{\sqrt{trS_{xx}^2\; trS_{yy}^2}} \tag{3.19}$$

et on voit que dist(Y, X) = 0 si et seulement si $RV_8 = 1$. Cette mesure est symétrique en x et en y mais pas invariante pour des transformations non-orthogonales. D'une façon plus

générale, si l'on munit X et Y des métriques définies positives M_x et M_y respectivement, la distance ci-dessus s'écrit

$$dist_{M_y,M_x}(Y, X) = \| \frac{Y' M_y Y}{\sqrt{tr(Y' M_y Y)^2}} - \frac{X' M_x X}{\sqrt{tr(X' M_x X)^2}} \| .$$

Comme précédemment on montre que $dist_{M_y,M_x}(Y, X) = \sqrt{2}\ \sqrt{1-RV^*_8}$

où $RV^*_8 = \dfrac{tr(S_{yx} M_x S_{xy} M_y)}{\sqrt{tr(S_{xx} M_x)^2\ tr(S_{yy} M_y)^2}}$.

On en déduit les relations suivantes

(i) si $M_x = I_q$ et $M_y = I_p$ alors $RV^*_8 = RV_8$

(ii) si $M_x = S_{xx}^{-1}$ et $M_y = S_{yy}^{-1}$ alors $RV^*_8 = \dfrac{1}{\sqrt{pq}} \sum_{i=1}^{p} \rho_i^2 = \sqrt{\dfrac{p}{q}} RV_5$

(iii) Si $M_x = S_{xx}^{-1}$ et $M_y = I_p$ alors $RV^*_8 = \dfrac{tr(S_{yx} S_{xx}^{-1} S_{xy})}{\sqrt{q\ tr\ S_{yy}^2}} = \dfrac{trS_{yy}}{\sqrt{q\ tr\ S_{yy}^2}} RV_5$

3.9. La mesure de Robert et Escoufier [17]

On obtient cette mesure de liaison en cherchant à maximiser RV_8 dans un contexte de régression multivariée. Si l'on cherche une transformation M: qxp telle que RV_8 $(Y, M'X)$ soit maximum sous la contrainte $M'\ S_{xx}\ M$ diagonale, on trouve que M doit satisfaire l'équation aux vecteurs propres généralisés $S_{yx}\ S_{xy}\ M - S_{xx}\ M\ \Lambda = 0$. Considérant alors la matrice H:pxp orthogonale qui diagonalise $S_{xy}\ S_{xx}^{-1}\ S_{yx}$ on vérifie aisément que $M' = HS_{yx}\ S_{xx}^{-1}$. Il en découle que les transformations M' et $M^{*'} = S_{yx}\ S_{xx}^{-1}$ fournissent la même valeur de RV_8 puisque M' et $M^{*'}$ se déduisent l'une de l'autre par la transformation orthogonale H. Cette valeur est

$$RV_9 = RV_8\ (Y, M^{*'}\ X) = \left[\frac{tr(S_{yx}\ S_{xx}^{-1}\ S_{xy})^2}{trS_{yy}^2} \right]^{\frac{1}{2}} = \left[\frac{trS_{yy}^2}{trS_{yy}^2} \right]^{\frac{1}{2}} . \qquad (3.20)$$

Il est évident que RV_9, tout comme RV_5, n'est pas symétrique en y et x ni invariante. On peut également l'écrire sous la forme

$$RV_9 = \left[\frac{tr(G'\Lambda G)^2}{tr(G'G)^2} \right]^{\frac{1}{2}} . \qquad (3.21)$$

3.10. Mesures de redondance et mesures d'association

Le Tableau 3.1 résume les mesures de liaison de cette section. Pour chacune on y indique ses auteurs ainsi que ses expressions en termes matriciels et en fonction des corrélations canoniques s'il y a lieu.

RV_5 et RV_9 sont des mesures de redondance puisqu'elles sont associées à la prédiction du vecteur $\underline{Y}$: px1 par le vecteur $\underline{X}$: qx1. Les autres mesures sont des mesures d'association. Elles indiquent le degré de dépendance entre les vecteurs $\underline{Y}$ et $\underline{X}$. Les propriétés de ces mesures de liaison sont les suivantes:

(i) $0 \leq RV_i \leq 1$, i=1,2,..., 9

(ii) si p=q=1, alors pour chaque i, RV_i devient le carré du coefficient de corrélation simple entre deux variables

(iii) si p=1, alors pour chaque $i \neq 8$, RV_i devient le carré du coefficient de corrélation multiple entre une variable et un vecteur

(iv) les mesures d'association, et non les mesures de redondance, sont symétriques en x et y

(v) pour chaque $i \neq 8$, RV_i est fonction des corrélations canoniques entre $\underline{Y}$ et $\underline{X}$

(vi) pour chaque $i \neq 5,8,9$, RV_i est invariant sous les transformations linéaires de $\underline{Y}$ et de $\underline{X}$.

On remarque que les mesures de redondance et celles d'association ne possèdent pas les même propriétés de symétrie et d'invariance et que RV_8 est une mesure d'association particulière. Robert et Escoufier [17] l'ont utilisée pour unifier différentes méthodes d'analyse multivariée.

Mesure de liaison	Proposée par	Expression en termes matriciels	Expression en fonction des corrélations canoniques
RV_1	Hotelling [8] Cramer [5]	$\frac{\lvert S_{\hat{y}\hat{y}}\rvert}{\lvert S_{yy}\rvert}$	$\prod_{i=1}^{p} \hat{\rho}_i^2$
RV_2	Hotelling [8] Roseboom [19]	$1- \frac{\lvert S_{ee}\rvert}{\lvert S_{yy}\rvert}$	$1- \prod_{i=1}^{p} (1-\hat{\rho}_i^2)$
RV_3	Cramer et Nicewander[6]	$\left(\frac{\lvert S_{\hat{y}\hat{y}}\rvert}{\lvert S_{yy}\rvert}\right)^{\frac{1}{p}}$	$\left(\prod_{i=1}^{p} \hat{\rho}_i^2\right)^{\frac{1}{p}}$
RV_4	Cramer et Nicewander[6]	$1-\left(\frac{\lvert S_{ee}\rvert}{\lvert S_{yy}\rvert}\right)^{\frac{1}{p}}$	$1-\left(\prod_{i=1}^{p} (1-\hat{\rho}_i^2)\right)^{\frac{1}{p}}$
RV_5	Stewart et Love [22]	$\frac{tr\ S_{\hat{y}\hat{y}}}{tr\ S_{yy}}$	$\frac{tr\ G'\Lambda G}{tr\ G'G}$
RV_6	Cramer et Nicewander[6]	$\frac{1}{p}\ tr(S_{yy}^{-1}\ S_{\hat{y}\hat{y}})$	$\frac{1}{p} \sum_{i=1}^{p} \hat{\rho}_i^2$
RV_7	Coxhead [4] et Shaffer et Gillo [20]	$\frac{tr(S_{ee}^{-1}\ S_{\hat{y}\hat{y}})}{tr(S_{ee}^{-1}\ S_{yy})}$	$\frac{\sum_{i=1}^{p} \hat{\rho}_i^2/(1-\hat{\rho}_i^2)}{\sum_{i=1}^{p} 1/(1-\hat{\rho}_i^2)}$
RV_8	Escoufier [7]	$\frac{tr(S_{yx}\ S_{xy})}{\left[tr\ S_{xx}^2\ tr\ S_{yy}^2\right]^{1/2}}$	———
RV_9	Robert et Escoufier [17]	$\left(\frac{tr\ S_{\hat{y}\hat{y}}^2}{tr\ S_{yy}^2}\right)^{1/2}$	$\left(\frac{tr(G'\Lambda G)^2}{tr(G'G)^2}\right)^{1/2}$

Tableau 3.1: Résumé des mesures de liaison de la Section 3.

4. RELATIONS ENTRE LES RV_i

Il existe une relation d'ordre partiel entre les mesures de liaison de la section précédente. Notons d'abord qu'il suit de (3.3) et (3.6)

$$RV_1 \leq \min_i \rho_i^2 \leq RV_3 . \tag{4.1}$$

Par (3.6) et (3.12) on a

$$RV_3 \leq RV_6 \tag{4.2}$$

puisque RV_3 est la moyenne géométrique des ρ_i^2 et que RV_6 en est la moyenne arithmétique. A partir de (3.12) et (3.7) on peut écrire.

$$\prod_{i=1}^{p} (1-\rho_i^2)^{\frac{1}{p}} \leq \frac{1}{p} \sum_{i=1}^{p} (1-\rho_i^2) = 1 - \frac{1}{p} \sum_{i=1}^{p} \rho_i^2 = 1-RV_6$$

d'où

$$RV_6 \leq RV_4 . \tag{4.3}$$

Par (3.18) et (3.7),

$$RV_4 \leq RV_7 \tag{4.4}$$

puisque $1-RV_7$ est la moyenne harmonique des quantités $(1-\rho_i^2)$ et que $1-RV_4$ en est la moyenne géométrique. De plus,

$$RV_7 \leq \max_i \rho_i^2 \tag{4.5}$$

car RV_7 est également une moyenne pondérée des ρ_i^2, par (3.7). Mais puisque $RV_2 \geq \max_i \rho_i^2$, il suit que

$$RV_7 \leq \max_i \rho_i^2 \leq RV_2 . \tag{4.6}$$

Finalement, en regroupant (4.1) à (4.6), nous obtenons

$$RV_1 \leq \min \rho_i^2 \leq RV_3 \leq RV_6 \leq RV_4 \leq RV_7 \leq \max_i \rho_i^2 \leq RV_2 . \tag{4.7}$$

D'autre part, par définition de RV_8 et RV_9, on a

$$RV_8 \leq RV_9 . \tag{4.8}$$

Nous allons maintenant montrer que

$$\frac{1}{\sqrt{p}}\, RV_5 \leq RV_9 \leq \sqrt{p}\, RV_5 \,. \tag{4.9}$$

Nous avons vu que $\sqrt{tr(E'E)}$ est une norme dont le produit scalaire associé est $tr(E'F)$. Par l'inégalité de Cauchy-Schwartz on a

$$(tr(E'F))^2 \leq tr(E'E)tr(F'F)$$

et si F = I et si E est symétrique, l'inégalité précédente devient

$$(trE)^2 \leq trE^2\, trI = p\, trE^2 \,. \tag{4.10}$$

En posant $E = S_{yy}$, (4.10) s'écrit

$$trS_{yy}^2 \geq \frac{1}{p}\,(trS_{yy})^2 \,. \tag{4.11}$$

D'autre part, nous avons

$$(trE)^2 = \left[\sum_{i=1}^{p} \lambda_i\right]^2 \geq \sum_{i=1}^{p} \lambda_i^2 = trE^2 \tag{4.12}$$

où $\lambda_1, \lambda_2, \ldots, \lambda_p$ sont les valeurs propres de E. Si l'on pose $E = S_{yx}S_{xx}^{-1}S_{xy}$ dans (4.12), il suit

$$tr(S_{yx}S_{xx}^{-1}S_{xy})^2 \leq (tr(S_{yx}S_{xx}^{-1}S_{xy}))^2 \,. \tag{4.13}$$

En combinant (4.11) et (4.13), on obtient

$$RV_9^2 \leq p\, RV_5^2 \,. \tag{4.14}$$

Si l'on pose $E = S_{yx}\, S_{xx}^{-1}\, S_{xy}$ dans (4.10) et $E = S_{yy}$ dans (4.12) on voit facilement que

$$\frac{1}{p}\, RV_5^2 \leq RV_9^2 \,. \tag{4.15}$$

Donc, la relation (4.7) ordonne les mesures d'association, (4.9) ordonne les mesures de redondance et (4.8) établit une relation entre RV_8 et RV_9.

5. TESTS D'INDEPENDANCE ENTRE DEUX VECTEURS ALEATOIRES

Supposons que la matrice de covariance (inconnue) du vecteur $\begin{bmatrix} Y \\ \underline{X} \end{bmatrix}$ soit

$$\begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}$$

et supposons également que les corrélations canoniques, au niveau de la population, soient $\rho_1^2 \geq \rho_2^2 \geq \cdots \geq \rho_p^2$. Alors pour chaque i=1,2,..., q, RV_i est un estimateur de ρV_i, où ρV_i est une mesure de liaison, au niveau de la population, qui s'exprime en fonction des ρ_i^2 ou de Σ_{xx}, Σ_{xy} et Σ_{yy} de la même façon que RV_i s'exprime en fonction de $\hat{\rho}_i^2$ ou de S_{xx}, S_{xy} et S_{yy}. / 9

Les deux vecteurs $\underline{Y}$ et $\underline{X}$ sont non corrélés si et seulement si $\Sigma_{xy} = 0$ ou si et seulement si $\rho_1 = \rho_2 = \cdots = \rho_p = 0$. Sous cette hypothèse, que nous noterons H_0, tous les ρV_i sont nuls.

Par la suite, afin de tester l'hypothèse H_0, nous supposerons que le vecteur $\begin{bmatrix} \underline{Y} \\ \underline{X} \end{bmatrix}$ est multinormal. Dans ce cas, H_0 signifie l'indépendance des deux vecteurs. Nous procéderons par un exemple.

Considérons le tableau de données présenté dans BMDP Statistical Software [2], p. 38. Il s'agit de 8 variables mesurées sur 188 patients: âge, taille, poids, prise de contraceptifs, taux de cholestérol, albumine, calcium et acide urique. On ne considérera pas ici la variable relative à la prise de contraceptifs. En outre, on omettra 7 patients pour lesquels certaines données sont manquantes. L'analyse portera donc sur 181 sujets et 7 variables.

Soient $\underline{Y}$ le vecteur composé des trois premières variables (les variables biologiques) et $\underline{X}$ le vecteur composé des quatre dernières (les variables biochimiques). Nous nous intéressons à la liaison entre les vecteurs $\underline{Y}$ et $\underline{X}$ et nous testerons l'hypothèse d'indépendance de ces deux vecteurs à partir des corrélations canoniques et de deux mesures de liaison de la Section 3.

On a n = 181, p = 3, q = 4 et on calcule

$$\bar{Y} = \begin{bmatrix} 33.49 \\ 64.49 \\ 131.20 \end{bmatrix} \quad , \quad \bar{X} = \begin{bmatrix} 234.81 \\ 4.12 \\ 9.97 \\ 4.75 \end{bmatrix}$$

puis

$$S_{yy} = \begin{bmatrix} 97.83 & 2.01 & 50.61 \\ 2.01 & 6.20 & 24.33 \\ 50.61 & 24.33 & 419.88 \end{bmatrix} \quad , \quad S_{xx} = \begin{bmatrix} 2007.69 & 1.21 & 5.50 & 14.86 \\ 1.21 & 0.13 & 0.08 & 0.02 \\ 5.50 & 0.08 & 0.22 & 0.09 \\ 14.86 & 0.02 & 0.09 & 1.27 \end{bmatrix}$$

$$S_{yx} = \begin{bmatrix} 162.43 & -0.26 & -0.02 & 2.38 \\ -2.42 & -0.01 & 0.16 & 0.31 \\ 107.99 & -1.75 & 0.59 & 6.76 \end{bmatrix}.$$

5.1. Test basé sur les corrélations canoniques

Les corrélations canoniques sont données par

$$\hat{\rho}_1 = .4645, \quad \hat{\rho}_2 = .3244 \quad \hat{\rho}_3 = .1225$$
$$\hat{\rho}_1^2 = .2158, \quad \hat{\rho}_2^2 = .1052, \quad \hat{\rho}_3^2 = .0150,$$

et on calcule

$$RV_1 = .0003, \quad RV_2 = .3088, \quad RV_3 = .0698$$
$$RV_4 = .1158, \quad RV_5 = .1654, \quad RV_6 = .1120$$
$$RV_7 = .1197, \quad RV_8 = .0433, \quad RV_9 = .1749$$

Si le vecteur $\begin{bmatrix} \underline{Y} \\ \underline{X} \end{bmatrix}$ est multinormal, le test du rapport des vraisemblances pour tester $H_k : \rho_{k+1} = \cdots = \rho_p = 0$, pour k = 0,1,..., p-1 est basé sur la statistique

$$W_k = \prod_{i=k+1}^{p} (1-\hat{\rho}_i^2) \tag{5.1}$$

et l'hypothèse H_k est rejetée au niveau α si

$$W_k^* = -[(n-1)-.5\ (p+q+1)]\ \ln W_k > c_\alpha$$

où c_α est la $100(1-\alpha)$ ième centile de la distribution $\chi^2_{(p-k)(q-k)}$ (voir Muirhead [15], Sect. 11.3.6). Ce test est un test asymptotique, pour n grand.

Ici on calcule

Corrélation canonique	W_k	W_k^*	Degrés de liberté	Niveau de signification
$\hat{\rho}_1 = .4645$	.6912	65.00	12	.000
$\hat{\rho}_2 = .3244$	.8814	22.22	6	.001

$\hat{\rho}_3 = .1235 \quad .9850 \quad 2.66 \quad 2 \quad .264$

et on doit conclure que seule la troisième corrélation canonique est nulle, que le rang de la matrice Σ_{yx} est 2 et que les vecteurs $\underline{Y}$ et $\underline{X}$ ne sont pas indépendants.

D'autres tests, basés également sur les corrélations canoniques, existent pour tester H_0. Voir par exemple Muirhead [15], sect. 4.2.8.

5.2. Test basé sur RV_5

Dans l'Appendice nous montrons que si le vecteur $\begin{bmatrix} \underline{Y} \\ \underline{X} \end{bmatrix}$ est multinormal et, sous H_0, la distribution de $RV_5/(1-RV_5)$ est donnée par $p[RV_5/(1-RV_5) \leq r] = p[W \leq 0]$ où W est distribué comme $\sum_{i=1}^{p(n-1)} \lambda_i W_i^2$ où les W_i sont iid et N(0,1), et où les λ_i sont les p valeurs propres de Σ_{yy} ayant chacune multiplicité q et les p valeurs propres de $-r\Sigma_{yy}$ ayant chacune multiplicité n-1-q.

On calculera donc la valeur r de $RV_5/(1-RV_5)$ puis la probabilité $p[RV_5/(1-RV_5) > r]$. Si cette probabilité est inférieure à .05, l'hypothèse H_0 sera rejetée. Si p=1, ce test devient le test F habituel de la régression linéaire multiple.

Pour calculer $p[RV_5/(1-RV_5) > r]$ en pratique on remplacera les λ_i par les $\hat{\lambda}_i$ obtenus à partir de S_{yy} à la place de Σ_{yy} et on utilisera l'algorithme de Imhof [9].

Dans le présent exemple nous calculons $RV_5 = .1654$ r = .1982 et $p[RV_5/(1-RV_5) > .1982] = 0.000$. On rejette donc l'hypothèse H_0, et les deux vecteurs $\underline{Y}$ et $\underline{X}$ ne sont pas indépendants.

5.3. Test basé sur RV_8

Dans Cléroux et Ducharme [3] il est montré que si $\begin{bmatrix} \underline{Y} \\ \underline{X} \end{bmatrix}$ est multinormale et, sous H_0, la distribution asymptotique (quand $n \to \infty$) de $n\,RV_8$ est celle de $U = \dfrac{1}{\sqrt{tr\Sigma_{yy}^2 tr\Sigma_{xx}^2}} \sum_{i=1}^{p}\sum_{j=1}^{q} \lambda_i \gamma_j U_{ij}^2$ où les λ_i sont les valeurs propres de Σ_{yy}, les γ_j sont celles de Σ_{xx} et les U_{ij} sont iid et N(0,1).

Un test asymptotique pour H_0 est donc le suivant: calculer la valeur u de $n\,RV_8$ puis la probabilité $p[U \geq u]$ en utilisant l'algorithme de Imhof [9]. L'hypothèse H_0 sera rejetée si cette probabilité est inférieure à .05. En pratique les λ_i et les γ_j seront remplacées respectivement par les valeurs propres $\hat{\lambda}_i$ de S_{yy} et $\hat{\gamma}_j$ de S_{xx}.

Ici on calcule RV_8 = .0433, u = 7.8373 et finalement p[U ≥ 7.8373] = .005. Comme précédemment on doit rejeter H_0 et conclure que $\underline{Y}$ et $\underline{X}$ ne sont pas indépendants.

La distribution asymptotique de RV_8 sous H_1: $\Sigma_{yx} \neq 0$ est obtenue dans Robert, Cléroux et Ranger [18] et permet de calculer, au besoin, la puissance du test précédent.

Les trois tests précédents ne sont pas équivalents pour plusieurs raisons:

(i) RV_5 et RV_8 ne sont pas des fonctions explicites des corrélations canoniques

(ii) le test basé sur les corrélations canoniques fournit à priori une information plus fine que les deux autres tests.

(iii) le test basé sur RV_5 est un test exact, les deux autres tests étant asymptotiques.

(iv) RV_8 est une mesure d'association qui est définie sans contexte particulier tandis que RV_5 est une mesure de redondance définie dans un contexte de régression.

D'autre part, les lois asymptotiques de W^*_k, RV_5 et RV_8 sont connues également sous l'hypothèse H_1:$\Sigma_{xy} \neq 0$ et les lois asymptotiques de RV_5 et RV_8 sont connues également sous la suite de contre hypothèses locales $\{H_{1n} : \Sigma_{xy} = \frac{A}{\sqrt{n}}\}$ où A : qxp est une matrice qui ne dépend pas explicitement de n. Il est donc possible d'effectuer des comparaisons numériques, par simulation de Monte Carlo, entre les puissances asymptotiques des tests précédents. Ce travail nécessitera d'énormes ressources informatiques. Il reste à faire.

APPENDICE: Distribution de $RV_5/(1-RV_5)$.

Supposons que le vecteur $\begin{bmatrix} Y \\ \underline{X} \end{bmatrix}$ soit multinormal avec matrice de covariance

$$\Sigma = \begin{bmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{bmatrix}$$

où Σ_{yy}: pxp et Σ_{xx}: qxq. Au niveau de l'échantillon de taille n, soit

$$S = \begin{bmatrix} S_{yy} & S_{yx} \\ S_{xy} & S_{xx} \end{bmatrix} = \frac{1}{n-1} \begin{bmatrix} A_{yy} & A_{yx} \\ A_{xy} & A_{xx} \end{bmatrix} = \frac{A}{n-1} .$$

Alors la distribution de la matrice A:(p+q)x(p+q) est la distribution de Wishart avec paramètres Σ et n-1, que l'on note $W_{p+q}(\Sigma, n-1)$.

Le théorème suivant est démontré dans Mardia, Kent et Bibby [12], p. 71. Il sera utile pour la suite.

Théorème:

a) $A_{yy.x} = A_{yy} - A_{yx} A_{xx}^{-1} A_{xy}$ possède la distribution $W_p(\Sigma_{yy.x}, n-1-q)$ où $\Sigma_{yy.x} = \Sigma_{yy} - \Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy}$ et $A_{yy.x}$ est indépendante de A_{xx} et de A_{yx}.

b) Sous H_0: $\Sigma_{yx} = 0$, $A_{yy} - A_{yy.x}$ possède la distribution $W_p(\Sigma_{yy}, q)$ et les matrices $A_{yx} A_{xx}^{-1} A_{xy}$, A_{xx} et $A_{yy.x}$ sont conjointement indépendantes.

A partir de (3.11) on peut écrire

$$RV_5 = \frac{tr(A_{yx}\, A_{xx}^{-1}\, A_{xy})}{trA_{yy.x} + tr(A_{yx}\, A_{xx}^{-1}\, A_{xy})}$$

d'où

$$\frac{RV_5}{1-RV_5} = \frac{tr(A_{yx}\, A_{xx}^{-1}\, A_{xy})}{trA_{yy.x}} .$$

Par le théorème précédent et sous H_0 on obtient

$$\frac{RV_5}{1-RV_5} = \frac{trW_p(\Sigma_{yy},\, q)}{trW_p(\Sigma_{yy},\, n-1-q)}$$

et le numérateur et le dénominateur sont indépendants.

On obtient maintenant la loi exacte de $RV_5/(1-RV_5)$. Par définition de la loi de Wishart, on peut écrire

$$W_p(\Sigma_{yy},\, q) = \sum_{i=1}^{q} Z_i Z_i' \text{ et } W_p(\Sigma_{yy},\, n-1-q) = \sum_{i=q+1}^{n-1} Z_i Z_i'$$

où $Z_1, Z_2, \ldots, Z_{n-1}$ sont iid et N $(0, \Sigma_{yy})$. Egalement,

$$V_1 = trW_p(\Sigma_{yy},\, q) = \sum_{i=1}^{q} Z_i' Z_i \text{ et } V_2 = trW_p(\Sigma_{yy},\, n-1-q) = \sum_{i=q+1}^{n-1} Z_i' Z_i .$$

L'objectif est donc de calculer la probabilité $p[\frac{V_1}{V_2} \le r] = p[V_1 - r\, V_2 \le 0]$. La décomposition spectrale de Σ_{yy} est $\Sigma_{yy} = SLS'$ où $L = diag(l_1, l_2, .., l_p)$ et où l_j, $j = 1,2,...,p$ sont les valeurs propres de Σ_{yy} supposée définie positive.

Pour $i = 1,2,...,n-1$, définissons $Y_i = L^{-\frac{1}{2}} S' Z_i$. Alors les Y_i sont indépendants, chacun de loi $N(0, I_p)$.

On peut écrire

$$V_1 = \sum_{i=1}^{q} Y_i' L\, Y_i = \sum_{i=1}^{q} \sum_{j=1}^{p} l_i\, Y_{ij}^2$$

et de la même façon $V_2 = \sum_{i=q+1}^{n-1} \sum_{j=1}^{p} l_i\, Y_{ij}^2$

où Y_{ij}, $j = 1,2,\ldots,p$ sont les composantes du vecteur Y_i . Finalement, $W = V_1 - r\, V_2$ est distribué comme $\sum_{i=1}^{p(n-1)} \lambda_i\, W_i^2$ où les W_i sont iid $N(0,1)$ et où les λ_i sont

(i) les p valeurs propres de Σ_{yy}, chacune ayant multiplicité q, et

(ii) les p valeurs propres de -r Σ_{yy}, chacune ayant multiplicité n-1-q.

Nous avons donc démontré le théorème suivant:

Théorème:

Si le vecteur $\begin{bmatrix} Y \\ \underline{X} \end{bmatrix}$ est multinormal et, sous H_0 , $p[RV_S/(1-RV_S) \leq r] = p[W \leq 0]$ où W est distribué comme $\sum_{i=1}^{p(n-1)} \lambda_i\, W_i^2$ où les W_i sont iid et N(0,1), et où $\lambda_1, \lambda_2, \ldots, \lambda_{p(n-1)}$ sont les p valeurs propes de Σ_{yy} ayant chacune multiplicité q et les p valeurs propres de $-r\Sigma_{yy}$ ayant chacune multiplicité n-1-q.

REMERCIEMENTS

Les auteurs remercient le Conseil de recherches en sciences naturelles et en génie du Canada ainsi que la fondation FCAR (Gouvernement du Québec) pour leur support financier. Ils remercient également l'éditeur et les rapporteurs pour leurs précieux conseils.

BIBLIOGRAPHIE

[1] Anderson, T.W.: An Introduction to Multivariate Statistical Analysis, 2nd ed., 1984, John Wiley, New York.

[2] BMDP Statistical Software, W.J. Dixon, Chief Editor, 1981, U. of Calif. Press, Berkeley.

[3] Cléroux, R. and G. Ducharme: Vector Correlation for Elliptical Distribution, 1986, Publ. 586, Département d'informatique et recherche opérationnelle, Université de Montréal.

[4] Coxhead, P.: Measuring the Relationships Between Two Sets of Variables, Brit. Jour. Math. Stat. Psycho., 1974, 27, 205-212.

[5] Cramer, E.M.: A Generalisation of Vector Correlation and its Relation to Canonical Correlation, Mult. Behav. Res., 1974, 9, 347-352.

[6] Cramer, E.M. and W.A. Nicewander: Some Symmetric, Invariant Measures of Multivariate Association, Psychometrika, 1979, 44, 43-54.

[7] Escoufier, Y.: Le traitement des variables vectorielles, Biometrics, 1973, 29, 751-760.

[8] Hotelling, H.: Relations Between Two Sets of Variables, Biometrika, 1936, 28, 321-377.

[9] Imhof, P.: Computing the Distribution of Quadratic Forms in Normal Variates, Biometrika, 1961, 48, 419-426.

[10] Johnson, N.L. and S. Kotz: Continuous Univariate Distributions, vol. 2, 1970, Hougton Mifflin, New York.

[11] Kshirsagar, A.M.: Correlation Between Two Vector Variables, 1969, Jour. Roy. Stat. Soc., Series B, 31, 477-485.

[12] Mardia, K.V., J.T. Kent and J.M. Bibby: Multivariate Analysis, 1979, Academic Press, London.

[13] Masuyama, M.: Correlation Between Tensor Quantities, Proc. Physico-Math. Soc. Japan, Series 3, 1939, 31, 638-647.

[14] Masuyama, M.: Correlation Coefficient Between Two Sets of Complex Vectors, Proc. Physico-Math. Soc. Japan, Series 3, 1941, 918-924.

[15] Muirhead, R.J.: Aspects of Multivariate Statistical Theory, 1982, John Wiley, New York.

[16] Ramsay, J.O., J. ten Berge and G. Styan: Matrix Correlation, Psychometrika, 1984, 49, 403-423.

[17] Robert, P. and Y. Escoufier: A Unifying Tool for Linear Multivariate Statistical Methods: the RV-Coefficient, Applied Stat., 1976, 25, 257-265.

[18] Robert, P., R. Cléroux and N. Ranger: Some Results on Vector Correlation, Comp. Stat. Data Anal., 1985, 3, 25-32.

[19] Roseboom, W.W.: Linear Correlation Between Sets of Variables, Psychometrika, 1965, 30, 57-71.

[20] Shaffer, J.P. and M.W. Gillo: A Multivariate Extension of the Correlation Ratio, Educ. Psycho. Measurements, 1974, 34, 521-524.

[21] Stephens, M.: Vector Correlation, Biometrika, 1979, 66, 41-48.

[22] Stewart, D. and W. Love: A General Canonical Correlation Index, Psycho. Bull., 1968, 70, 160-163.