REVUE DE STATISTIQUE APPLIQUÉE

AZIZ LAZRAQ
ROBERT CLÉROUX
HENK A. L. KIERS

Mesures de liaison vectorielle et généralisation de l'analyse canonique

Revue de statistique appliquée, tome 40, n° 1 (1992), p. 23-35
<http://www.numdam.org/item?id=RSA_1992__40_1_23_0>

Rev. Statistique Appliquée, 1992, **XXXIX** (1), 23-35

MESURES DE LIAISON VECTORIELLE ET GÉNÉRALISATION DE L'ANALYSE CANONIQUE[1]

Aziz LAZRAQ
Ecole Nationale de l'Industrie Minérale, Rabat

Robert CLÉROUX
Université de Montréal

Henk A.L. KIERS
Université de Groningen

SOMMAIRE

Plusieurs généralisations de l'analyse canonique furent proposées pour plus de deux populations. L'une d'elles est due à Carroll (1968). Dans cet article nous présentons l'approche de Carroll dans le cas vectoriel en utilisant le coefficient de corrélation vectorielle de Lingoes et Schönemann (1974) puis nous utilisons de la même façon le coefficient de Escoufier (1973).

Mots-clés : *Analyse canonique généralisée, Corrélation vectorielle, Mesures de liaison.*

SUMMARY

Many authors have proposed generalizations of classical canonical analysis to more than two groups. One method is due to Carroll (1968). In this paper we present his approach in a vectorial context using the vectorial correlation coefficient introduced by Lingoes and Schönemann (1974) and then, we use Escoufier's coefficient (1973) in a similar manner.

Key-words : *Generalized canonical correlation, Vectorial correlation, Measures of association.*

Introduction

L'analyse canonique classique consiste à trouver une combinaison linéaire des p variables d'un vecteur ainsi qu'une autre combinaison linéaire des q variables

[1] Cette recherche fut entreprise alors que les auteurs se trouvaient à l'Unité de Biométrie, Université de Montpellier. Ils remercient le Professeur Escoufier de l'avoir rendue possible. Les auteurs remercient le lecteur anonyme de cet article pour ses précieux commentaires.

d'un second vecteur de telle sorte que le coefficient de corrélation simple entre les deux soit maximum. Plusieurs généralisations furent proposées pour $n > 2$ groupes : voir par exemple Kettenring (1971) et Carroll (1968). Carroll obtient une combinaison linéaire des variables pour chaque groupe, ainsi qu'une variable compromis Z, de telle sorte qu'une moyenne pondérée des carrés des coefficients de corrélation simple entre Z et les n combinaisons linéaires soit maximum. Plusieurs dimensions orthogonales peuvent être obtenues de la sorte. Dans cet article nous présentons l'approche de Carroll dans un contexte vectoriel. Nous utilisons les coefficients de corrélation vectorielle de Lingoes et Schönemann (1974) et de Escoufier (1973). Nous obtenons alors un ensemble de variables canoniques pour chacun des n groupes ainsi qu'un ensemble de variables compromis. Nous comparons ces résultats avec ceux obtenus en utilisant l'approche de Carroll.

Mais auparavant nous rappelons brièvement les mesures de liaison de Lingoes et Schönemann (RLS), l'indice de redondance RI de Stewart et Love (1968), la mesure de Escoufier (RV) et le coefficient RV_{reg} de Robert et Escoufier (1976), et établissons un parallèle entre les couples $(RLS,\ RI)$ et $(RV,\ RV_{reg})$.

En fait, cet article se compose de deux parties : les Sections 1 à 7 traitent des mesures de liaison et les Sections 8 à 11 généralisent l'analyse canonique au contexte vectoriel. La première partie est nécessaire à la bonne compréhension de la seconde.

2. Contexte de travail et notation

Considérons deux vecteurs de variables $X^{(1)} : p \times 1$ et $X^{(2)} : q \times 1$ mesurées sur les mêmes n sujets pour obtenir un échantillon $\begin{pmatrix} X_\alpha^{(1)} \\ X_\alpha^{(2)} \end{pmatrix}$ $\alpha = 1, 2, \ldots, n$ du vecteur $\begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix}$.

Posons

$$\overline{X} = \begin{pmatrix} \overline{X}^{(1)} \\ \overline{X}^{(2)} \end{pmatrix} \text{ et } S = \begin{pmatrix} S_{11} & S_{12} \\ X_{21} & S_{22} \end{pmatrix},$$

le vecteur des moyennes et la matrice des covariances où

$$\overline{X}^{(i)} = \frac{1}{n} \sum_{\alpha=1}^{n} X_\alpha^{(i)} \text{ pour } i = 1,\ 2$$

$$S_{ij} = \frac{1}{n-1} \sum_{\alpha=1}^{n} (X_\alpha^{(i)} - \overline{X}^{(i)})(X_\alpha^{(j)} - \overline{X}^{(j)})' \text{ pour } i, j = 1, 2$$

et où, pour toute matrice A, A' dénote sa transposée. A partir de cet échantillon de taille n on forme les matrices de données.

$$Y_1 = (X_1^{(1)} - \overline{X}^{(1)}, X_2^{(1)} - \overline{X}^{(1)}, \ldots, X_n^{(1)} - \overline{X}^{(1)}) : p \times n$$
$$\text{et } Y_2 = (X_1^{(2)} - \overline{X}^{(2)}, X_2^{(2)} - \overline{X}^{(2)}, \ldots, X_n^{(2)} - \overline{X}^{(2)}) : q \times n$$

Rappelons que pour toute matrice E, la fonction $\| E \| = \sqrt{tr\ (E'E)}$ est une norme et que la distance entre deux matrices E et F de même taille est

$$\text{dist } (E, F) = \| E - F \| . \tag{2.1}$$

3. La mesure de liaison de Lingoes et Schönemann

Si $p = q$, les matrices Y_1 et Y_2 ont même taille. Si $p \neq q$, disons que $q < p$, on ajoute une colonne de $p-q$ zéros à chaque vecteur $X_\alpha^{(2)} - \overline{X}^{(2)}$, $\alpha = 1, 2, \ldots, n$. Alors la nouvelle matrice Y_2 devient de même taille que Y_1. Cette taille commune est max $(p, q) \times n$.

La qualité de l'approximation de Y_1 par Y_2 est mesurée en posant $E = Y_1$ et $F = Y_2$ dans (2.1) :

$$\begin{aligned} \text{dist}^2(Y_1, Y_2) &= \| Y_1 - Y_2 \|^2 = tr(Y_1 - Y_2)(Y_1 - Y_2)' \\ &= \sum_{\alpha=1}^{n} \| Y_{1\alpha} - Y_{2\alpha} \|^2 \end{aligned} \tag{3.1}$$
$$\text{où } Y_{1\alpha} = X_\alpha^{(1)} - \overline{X}^{(1)} \text{ et } Y_{2\alpha} = X_\alpha^{(2)} - \overline{X}^{(2)},\ \alpha = 1, 2, \ldots, n.$$

Si l'on cherche à transformer Y_2 linéairement par une transformation orthogonale T de sorte que la distance (3.1) soit minimale, c'est-à-dire

$$\begin{aligned} \min_{T(\perp)} \text{dist}^2(Y_1, TY_2) &= tr(Y_1 - TY_2)(Y_1 - TY_2)' \\ &= tr\ Y_1Y_1' + tr\ Y_2Y_2' - 2tr\ TY_2Y_1', \end{aligned} \tag{3.2}$$

le problème revient à trouver $\max_{T(\perp)} tr\ TY_2Y_1'$ où $T(\perp)$ dénote l'ensemble des matrices orthogonales. On utilisera le lemme suivant énoncé sans preuve (voir par exemple Green (1969)).

Lemme : Soit $A : r \times r$ une matrice carrée et $A = U\Lambda V'$ sa décomposition singulière où U et V sont orthogonales et $\Lambda = \text{diag } (\lambda_1, \lambda_2, \ldots, \lambda_r)$ avec $\lambda_i \geq 0$ pour $i = 1, 2, ..., r$. Alors, pour toute matrice T orthogonale, $tr\ TA \leq tr(A'A)^{1/2}$ et l'égalité tient si $T = VU'$.

Avec $A = Y_2\ Y_1'$ on obtient $\max_{T(\perp)} tr\ TY_2Y_1' = tr\ (Y_1Y_2'Y_2Y_1')^{1/2}$ et (3.2) devient

$$\min_{T(\perp)} \text{dist}^2(Y_1, TY_2) = tr\ Y_1Y_1' + tr\ Y_2Y_2' - 2tr\ (Y_1Y_2'Y_2Y_1')^{1/2}. \tag{3.3}$$

Si, de plus, les matrices Y_1 et Y_2 sont remplacées par

$$Y_1^* = \frac{Y_1}{\sqrt{tr\ Y_1'Y_1}} \text{ et } Y_2^* = \frac{Y_2}{\sqrt{tr\ Y_2'Y_2}} \text{ dans (3.1),}$$

on obtient, pour cette même transformation T, avec

$$S_{11} = \frac{Y_1Y_1'}{n-1},\ S_{22} = \frac{Y_2Y_2'}{n-1} \text{ et } S_{12} = \frac{Y_1Y_2'}{n-1},$$

$$\underset{T}{\text{Min}} \parallel \frac{Y_1}{\sqrt{tr\ Y_1'Y_1}} - \frac{TY_2}{\sqrt{tr\ Y_2'Y_2}} \parallel = \sqrt{2}\sqrt{1 - RLS} \tag{3.4}$$

où

$$RLS = RLS\ (Y_1, Y_2) = \frac{tr\ (S_{12}S_{21})^{1/2}}{\sqrt{tr\ S_{11}\ tr\ S_{22}}} \tag{3.5}$$

est la mesure de liaison de Lingoes et Schönemann (1974) entre deux tableaux de données. De (3.4) on a $RLS = 1$ si et seulement si il existe une rotation T qui fait complètement coïncider les deux configurations de n points.

Si A est non-singulière alors $TA = (A'A)^{1/2}$ possède une solution unique $T = (A'A)^{-1/2}A' = (Y_1Y_2'Y_2Y_1')^{-1/2}Y_1Y_2'$.

La rotation T qui transforme Y_2 en TY_2 est appelée rotation procruste de Y_2 relativement à Y_1.

Les propriétés de RLS sont les suivantes : (i) si $p = q = 1$ alors $RLS = |r|$ où r est le coefficient de corrélation simple empirique entre les variables X_1 et X_2, (ii) $0 \leq RLS \leq 1$ et $RLS = 0$ si et seulement si $S_{12} = 0$, (iii) $RLS\ (Y_1, Y_2) = RLS\ (Y_2, Y_1)$, (iv) si $H : p \times p$ est telle que $H'H = kI$ où $k > 0$ est un scalaire et I la matrice identité, alors $RLS\ (HY_1,\ Y_2) = RLS\ (Y_1,\ Y_2)$.

4. La mesure de liaison de Stewart et Love

On se situe dans le contexte de la régression linéaire multivariée : on cherche une transformation linéaire $M : q \times p$ telle que $RLS\ (Y_1,\ M'\ Y_2)$ soit maximum. On montre que $M' = S_{12}\ S_{22}^{-1}$ (voir Appendice 1) et pour cette valeur de M' on a $(Y_1 - M'\ Y_2)Y_2' = 0$. En reportant cette valeur de M' dans l'expression de $RLS\ (Y_1,\ M'\ Y_2)$ on obtient

$$RLS^2\ (Y_1,\ M'\ Y_2) = \frac{tr\ S_{12}S_{22}^{-1}S_{21}}{tr\ S_{11}} = RI\ (Y_1,\ Y_2) = RI, \tag{4.1}$$

l'indice de redondance de Stewart et Love (1968). La matrice des covariances résiduelles après la prédiction de $X^{(1)}$ par $X^{(2)}$, ou de Y_1 par Y_2, est donnée par $S_{11.2} = S_{11} - S_{12}\ S_{22}^{-1}\ S_{21}$. En prenant les traces il vient $tr\ S_{11.2} = (tr\ S_{11})(1 - RI)$ et, en multipliant par $n - 1$,

$$|Y_1 - M'\ Y_2|^2 = |Y_1|^2\ (1 - RI). \tag{4.2}$$

La mesure $tr\ S_{11.2}$ fut proposée par Rao (1965) pour évaluer la qualité de la prédiction de $X^{(1)}$ par $M'X^{(2)}$ en A.C.P.V.I. (analyse en composantes principales par rapport à des variables instrumentales). La mesure RI a été utilisée par Lazraq et Cléroux (1988B) dans une procédure pas à pas (de type stepwise) de sélection de variables en régression linéaire multivariée ou en A.C.P.V.I.

5. La mesure de liaison d'Escoufier

Cette mesure est également basée sur une certaine distance entre Y_1 et Y_2. Si dans (2.1) on pose

$$E = \frac{Y_1'Y_1}{\sqrt{tr\ (Y_1'Y_1)^2}} \text{ et } F = \frac{Y_2'Y_2}{\sqrt{tr\ (Y_2'Y_2)^2}},$$

une distance induite entre Y_1 et Y_2 est

$$d(Y_1, Y_2) = dist(E, F) = \left| \frac{Y_1'Y_1}{\sqrt{tr\ (Y_1'Y_1)^2}} - \frac{Y_2'Y_2}{\sqrt{tr\ (Y_2'Y_2)}} \right| \tag{5.1}$$

Il suit (voir Robert et Escoufier (1976)) que

$$d(Y_1, Y_2) = \sqrt{2}\sqrt{1 - RV} \tag{5.2}$$

$$\text{où } RV = RV(Y_1, Y_2) = \frac{tr\ S_{12}\ S_{21}}{\sqrt{tr\ S_{11}^2\ tr\ S_{22}^2}} \tag{5.3}$$

est la mesure de liaison de Escoufier (1973) entre deux tableaux de données. De (5.2) on a $RV = 1$ si et seulement si les positions relatives des n points dans $\mathbb{R}^p$ et des n points dans $\mathbb{R}^q$ sont semblables (*i.e.* se déduisent l'une de l'autre par un déplacement).

Les propriétés de RV (voir Escoufier (1973)) sont les suivantes : (i) si $p = q = 1$ alors $RV = r^2$ où r est le coefficient de corrélation simple empirique entre les variables X_1 et X_2, (ii) $0 \leq RV \leq 1$ et $RV = 0$ si et seulement si $S_{12} = 0$, (iii) $RV\ (Y_1, Y_2) = RV\ (Y_2, Y_1)$, (iv) si $H : p \times p$ est telle que $H'H = kI$ où $k > 0$ est un scalaire et I la matrice identité, alors $RV\ (HY_1,\ Y_2) = RV\ (Y_1,\ Y_2)$.

6. La mesure de Robert et Escoufier

Par analogie à la Section 4, on se situe dans le contexte de régression linéaire multivariée et on cherche une transformation linéaire $M : q \times p$ telle que $RV\ (Y_1, M'Y_2)$ soit maximum. On montre que $M' = S_{12}\ S_{22}^{-1}$ (voir Robert et Escoufier (1976)), et pour cette valeur de M' on a $(Y_1 - M'Y_2)Y_2' = 0$. En reportant cette valeur de M' dans (5.3) on obtient

$$RV\ (Y_1,\ M'Y_2) = \left(\frac{tr\ (S_{12}S_{22}^{-1}S_{21})^2}{tr\ S_{11}^2}\right)^{1/2} = RV_{reg}\ (Y_1,\ Y_2) = RV_{reg}. \quad (6.1)$$

Il est facile de voir que

$$|Y_1'Y_1 - Y_2'MM'Y_2|^2 = |Y_1'Y_1|^2\ (1 - RV_{reg}^2). \quad (6.2)$$

7. Comparaison entre ces quatre mesures de liaison

On a donc $RLS^2 \leq RI$ et $RV \leq RV_{reg}$. On montre également les relations

$$\frac{1}{\sqrt{p}} RI \leq RV_{reg} \leq \sqrt{p} RI \quad \text{(voir Lazraq et Cléroux (1988A))}$$

$$\frac{1}{\sqrt{pq}} RLS^2 \leq RV \leq \sqrt{pq}\ RLS^2. \quad \text{(voir Appendice 2).}$$

Le Tableau 7.1 résume les résultats obtenus sur les mesures de liaison RLS, RI, RV et RV_{reg}.

8. Rappel de l'analyse canonique généralisée de Carroll

Carroll (1968) généralise l'analyse canonique à $n > 2$ groupes en cherchant une combinaison linéaire des variables pour chaque groupe ainsi qu'une variable compromis la plus corrélée possible avec l'ensemble des n combinaisons linéaires. Au niveau de l'échantillon le problème s'énonce comme suit.

On dispose de n matrices de données $X_1,\ X_2, ...,\ X_n$ où $X_i : m_i \times k$ est formée des k observations prises sur chacune des m_i variables, $i = 1, 2, ..., n$. Chaque matrice X_i est supposée centrée (la somme de ses rangées est un vecteur de zéros) et de plein rang $m_i, i = 1, 2, ..., n$. On cherche un vecteur $Z : 1 \times k$ formé de k observations prises sur une variable compromis supposée centrée et n transformations linéaires $A_iX_i,\ A_i : 1 \times m_i$, telles que la quantité

$$R^2 = \sum_{i=1}^{n} \omega_i r^2(Z, A_iX_i) \quad (8.1)$$

TABLEAU 7.1
Comparaison entre les couples (RLS, RI) et (RV, RV_{reg})

Distance $RLS = \dfrac{tr\ (S_{12}\ S_{21})^{1/2}}{\sqrt{tr\ S_{11}\ tr\ S_{22}}}$ $\min_{T(\perp)} \left\| \dfrac{Y_1}{\sqrt{tr\ Y_1'Y_1}} - \dfrac{TY_2}{\sqrt{tr\ Y_2'Y_2}} \right\|$ $= \sqrt{2}\sqrt{1 - RLS}$	Distance $RV = \dfrac{tr\ S_{12}\ S_{21}}{\sqrt{tr\ S_{11}^2\ tr\ S_{22}^2}}$ $\left\| \dfrac{Y_1'Y_1}{\sqrt{tr\ (Y_1'Y_1)^2}} - \dfrac{Y_2'Y_2}{\sqrt{tr\ (Y_2'Y_2)^2}} \right\|$ $= \sqrt{2}\sqrt{1 - RV}$
Régression multivariée $M' = S_{12}S_{22}^{-1}$ $RI = \dfrac{tr\ S_{12}S_{22}^{-1}S_{21}}{tr\ S_{11}}$ $\lvert Y_1 - M'Y_2\rvert^2 = \lvert Y_1\rvert^2(1 - RI)$	Régression multivariée $M' = S_{12}S_{22}^{-1}$ $RV_{reg} = \left(\dfrac{tr\ (S_{12}S_{22}^{-1}S_{21})^2}{tr\ S_{11}^2}\right)^{1/2}$ $\lvert Y_1'Y_1 - Y_2'MM'Y_2\rvert = \lvert Y_1'Y_1\rvert^2(1 - RV_{reg}^2)$
Relations $RLS^2 \leq RI$	Relations $RV \leq RV_{reg}$
$\dfrac{1}{\sqrt{pq}} RLS^2 \leq RV \leq \sqrt{pq}\, RLS^2$ $\dfrac{1}{\sqrt{p}} RI \leq RV_{reg} \leq \sqrt{p}\, RI$	

soit maximale où ω_i est un poids associé à X_i et où $r\ (Z,\ A_iX_i)$ est le coefficient de corrélation simple empirique calculé à partir des composantes de Z et de A_iX_i. Pour Z fixé, la valeur maximale de $r^2\ (Z,\ A_iX_i)$ est atteinte quand $A_i = ZX_i'(X_iX_i')^{-1}$ et est donnée par

$$\max_{A_i} r^2(Z,\ A_iX_i) = \frac{ZX_i'(X_iX_i')^{-1}X_iZ'}{ZZ'} \tag{8.2}$$

En reportant (8.2) dans (8.1) on obtient

$$\begin{aligned} R^2 &= \frac{1}{ZZ'} \sum_{i=1}^{n} \omega_i ZX_i'(X_iX_i')^{-1}X_iZ' \\ &= \frac{ZQZ'}{ZZ'} \end{aligned} \tag{8.3}$$

où

$$Q = \sum_{i=1}^{n} \omega_i X_i' (X_i X_i')^{-1} X_i : k \times k \tag{8.4}$$

est au moins semi-définie positive. Dans une seconde étape on cherche à maximiser (8.3) par rapport à Z. C'est un problème classique de valeurs et de vecteurs propres de Q. Pour obtenir plus d'une dimension on procède comme suit : la seconde dimension est la meilleure dimension orthogonale à la première, la troisième est la meilleure orthogonale aux deux premières et ainsi de suite. La solution, dans r dimensions, est la matrice Z qui contient les r premiers vecteurs propres de la matrice Q.

9. Analyse canonique généralisée au cas vectoriel via (RLS, RI)

Le contexte est le même que dans la section précédente sauf que l'on cherche l combinaisons linéaires des variables de chaque groupe ainsi que l variables compromis les plus liées, au sens de la mesure RLS, avec l'ensemble des groupes de l combinaisons linéaires. Au niveau de l'échantillon on cherche une matrice $Z : l \times k$ formée de k observations prises sur chacune des l variables compromis et n groupes de combinaisons linéaires $A_i X_i$ où $A_i : l \times m_i$, $i = 1, 2, ..., n$ telles que la quantité

$$S^2 = \sum_{i=1}^{n} \omega_i RLS^2 \ (Z, \ A_i X_i) \tag{9.1}$$

soit maximale. Comme Carroll, nous travaillons sous la contrainte $ZZ' = I$. Pour Z fixé, la valeur maximale de $RLS^2 \ (Z, \ A_i X_i)$ est atteinte quand $A_i = ZX_i' \ (X_i X_i')^{-1}$ et est donnée par

$$RI = \frac{tr \ ZX_i' \ (X_i X_i')^{-1} X_i Z'}{tr \ ZZ'} = \frac{tr \ S_{zi} S_{ii}^{-1} S_{iz}}{tr \ S_{zz}} \tag{9.2}$$

où

$$(n-1) \ S_{zi} = ZX_i', \ (n-1) \ S_{ii} = X_i X_i', \ (n-1) \ S_{zz} = ZZ' = I \tag{9.3}$$

sont des matrices de covariance empiriques. La formule (9.1) peut s'écrire sous la forme

$$S^2 = \frac{tr \ ZQZ'}{tr \ ZZ'} = \frac{tr \ ZQZ'}{l} \tag{9.4}$$

où Q est donné par (8.4).

Dans une seconde étape on cherche à maximiser (9.4) par rapport à Z. Il s'agit donc de prendre pour Z les vecteurs propres de Q. On retrouve la solution de Carroll.

La solution de ce problème d'analyse canonique généralisée au cas vectoriel est donc identique à celle obtenue par la méthode de Carroll (1968). Le passage au cas vectoriel en utilisant le couple (RLS, RI) n'apporte aucun éclairage nouveau. Ce n'est pas le cas, cependant, si on utilise le couple $(RV,\ RV_{reg})$.

10. Analyse canonique généralisée au cas vectoriel via (RV, RV_{reg})

Le contexte est le même qu'à la section précédente sauf que l'on cherche à maximiser, par rapport à $Z : l \times k$ et à $A_i : l \times m_i,\ i = 1, 2, ..., n$ la quantité

$$T^2 = \sum_{i=1}^{n} \omega_i RV^2\ (Z,\ A_i X_i) \tag{10.1}$$

sous la contrainte $ZZ' = I$. Pour Z fixé, la valeur maximale de $RV^2\ (Z,\ A_i X_i)$ est atteinte quand $A_i = ZX_i'\ (X_i X_i')^{-1}$ et est donnée par

$$RV_{reg}^2 = \frac{tr\ (ZX_i'\ (X_i X_i')^{-1} X_i Z')^2}{tr\ (Z'Z)^2} = \frac{tr\ (S_{zi} S_{ii}^{-1} S_{iz})^2}{tr\ S_{zz}^2} \tag{10.2}$$

où S_{zi}, S_{ii} et S_{zz} sont données par (9.3). Le numérateur de (10.2) peut s'écrire

$$\begin{aligned} &tr\ ZX_i'(X_i X_i')^{-1} X_i Z' ZX_i'(X_i X_i')^{-1} X_i Z' \\ &\quad = tr\ (ZP_i Z')^2 = tr\ (P_i H_z)^2 \end{aligned} \tag{10.3}$$

où $H_z = Z'Z : k \times k$ est une matrice de produits scalaires et $P_i = X_i'\ (X_i X_i')^{-1}\ X_i : k \times k$ est semi-définie positive. Le dénominateur de (10.2) peut s'écrire

$$tr\ (ZZ')^2 = tr\ ZZ'ZZ' = tr\ I = l \tag{10.4}$$

Par (10.2), (10.3) et (10.4), la fonction à maximiser, par rapport à H_z, sous la contrainte $tr\ ZZ' = I$ est

$$\sum_{i=1}^{n} \omega_i tr\ (ZP_i Z')^2 \tag{10.5}$$

Cette fonction peut être maximisée de façon itérative en prenant, comme mise à jour de Z, la matrice UV' formée à partir de la décomposition singulière $\sum_{i=1}^{n} P_i Z' Z P_i Z' = UDV'$ (voir Kiers, Cléroux et Ten Berge (1991)). Cette procédure assure une convergence vers un maximum au moins local de (10.5). Un algorithme numérique a été codé en PCMATLAB pour calculer la solution Z obtenue par la méthode de Carroll (1968) ainsi que la solution $Z_1,\ Z_2, ..., Z_p$ de cette section.

11. Un exemple

Les données utilisées dans cet exemple se trouvent dans Johnson et Wichern (1982, p. 287). On a mesuré 7 variables sur 35 mouches de types leptoconops torrens. Ces variables sont la longueur de l'aile, la largeur de l'aile, la longueur de la troisième palpe, la largeur de la troisième palpe, la longueur de la quatrième palpe, la longueur du 12e segment de l'antenne et la longueur du 13e segment de l'antenne.

On standardise les données puis on forme trois matrices de données : $X_1 : 2 \times 35$ à partir des deux premières variables (données sur l'aile), $X_2 : 3 \times 35$ à partir des trois suivantes (données sur les palpes) et $X_3 : 2 \times 35$ à partir des deux dernières (données sur l'antenne). On a ici $n = 3$ tableaux, $m_1 = 2$, $m_2 = 3$, $m_3 = 2$ et $k = 35$ observations et on effectue une analyse canonique généralisée de ces trois tableaux, au sens de Carroll (1968) ou du couple $(RLS,\ RI)$ et au sens du couple $(RV,\ RV_{reg})$, en ne retenant que deux dimensions.

Nous avons pris $\omega_i = 1$, $i = 1, 2, 3$ et pour des raisons évidentes d'espace, seulement les résultats pour $k = 10$ (les 10 premières mouches) seront mentionnés.

Les résultats de l'analyse au sens de Carroll (1968) sont les suivants. Les valeurs propres de la matrice Q sont 2.42, 1.94, 1.22, 0.77, 0.39, 0.21 et 0.06. Nous nous limiterons donc à $l = 2$. Le maximum de la somme des $(RLS)^2$ est 2.18. La matrice Z compromis qui représente au mieux les trois matrices de données, au sens de Carroll et dans deux dimensions, est présentée au Tableau 1.

TABLEAU 1
Dimensions compromis de l'analyse canonique généralisée
basée sur les couples (RLS, RI) et (RV, RV_{reg})

	$(RLS,\ RI)$		$(RV,\ RV_{reg})$	
	Dim 1	Dim 2	Dim 1	Dim 2
1	.15	.79	.13	.74
2	.80	−.16	.83	−.16
3	−.44	−.25	−.43	−.30
4	−.16	.25	−.10	.12
5	−.11	−.24	−.06	−.25
6	−.17	.11	−.20	.21
7	−.13	.03	−.15	.07
8	−.17	−.01	−.16	.06
9	.08	−.18	.08	−.04
10	.14	−.35	.07	−.46

Pour l'analyse au sens du couple (RV, RV_{reg}) nous avons également choisi $l = 2$ pour fins de comparaison avec la méthode de Carroll. Le maximum de la somme des $(RV)^2$ est 1.75. La matrice compromis qui représente au mieux les trois matrices de données, au sens du couple (RV, RV_{reg}) et dans deux dimensions,

est également présentée au Tableau 1. Les résultats des deux analyses diffèrent peu sauf peut-être pour quelques composantes dans la seconde dimension.

Appendice I

On montre que la matrice M' qui maximise $RLS\ (Y_1,\ M'Y_2)$ est $M' = S_{12}S_{22}^{-1}$. On peut écrire

$$RLS(Y_1,\ M'Y_2) = \max_T \frac{tr\ Y_1'TM'Y_2}{(tr\ Y_1Y_1')^{1/2}(tr\ M'Y_2Y_2'M)^{1/2}} \tag{I.1}$$

où T est contrainte à l'orthonormalité. Par l'inégalité de Cauchy-Schwartz on a

$$\begin{aligned} tr\ Y_1'TM'Y_2 &= tr\ Y_2'\ (Y_2Y_2')^{-1}Y_2Y_1'TM'Y_2 \\ &\leq (tr\ Y_1Y_2'\ (Y_2Y_2')^{-1}Y_2Y_1')^{1/2}(tr\ TM'Y_2Y_2'M'T')^{1/2} \\ &= (tr\ S_{12}S_{22}^{-1}S_{21})^{1/2}(tr\ M'\ S_{22}\ M)^{1/2}(n-1). \end{aligned} \tag{I.2}$$

En reportant (I.2) dans (I.1) il suit

$$\begin{aligned} RLS\ (Y_1,\ M'Y_2) &\leq \frac{(tr\ S_{12}S_{22}^{-1}S_{21})^{1/2}(tr\ M'S_{22}M)^{1/2}}{(tr\ S_{11})^{1/2}(tr\ M'S_{22}M)^{1/2}} \\ &= \frac{(tr\ S_{12}S_{22}^{-1}S_{21})^{1/2}}{(tr\ S_{11})^{1/2}}, \end{aligned} \tag{I.3}$$

fournissant ainsi une borne supérieure pour $RLS\ (Y_1, M'Y_2)$ qui ne dépend pas de M ni de T. Pour montrer que cette borne supérieure est atteinte quand $M' = S_{12}S_{22}^{-1}$, il suffit de substituer $M' = S_{12}S_{22}^{-1}$ dans (I.1) :

$$\begin{aligned} RLS\ (Y_1, S_{12}S_{22}^{-1}Y_2) &= \max_T \frac{tr\ Y_1'TS_{12}S_{22}^{-1}Y_2}{(tr\ S_{11})^{1/2}(tr\ S_{12}S_{22}^{-1}S_{21})^{1/2}} \\ &= \max_T \frac{tr\ TS_{12}S_{22}^{-1}S_{21}}{(tr\ S_{11})^{1/2}(tr\ S_{12}S_{22}^{-1}S_{21})^{1/2}} \end{aligned} \tag{I.4}$$

Le maximum est atteint pour $T = I$ puisqu'alors on obtient la borne supérieure fournie par (I.3). On obtient finalement

$$RLS\ (Y_1,\ S_{12}S_{22}^{-1}Y_2) = \frac{(tr\ S_{12}S_{22}^{-1}S_{21})^{1/2}}{(tr\ S_{11})^{1/2}} = \sqrt{RI(Y_1,\ Y_2)}\ .$$

Appendice II

On montre que $\frac{1}{\sqrt{pq}} RLS^2 \leq RV \leq \sqrt{pq} RLS^2$.

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

$$(tr\ E'F)^2 \leq (tr\ E'E)(tr\ F'F). \tag{II.1}$$

Si $F = I_l$ et E est symétrique, (II.1) devient

$$(tr\ E)^2 \leq l\ tr\ E^2. \tag{II.2}$$

En posant $E = S_{11}$ et $E = S_{22}$ respectivement dans (II.2), il vient

$$\begin{aligned} tr\ S_{11}^2 &\geq \frac{1}{p}(tr\ S_{11})^2 \\ tr\ S_{22}^2 &\geq \frac{1}{q}(tr\ S_{22})^2. \end{aligned} \tag{II.3}$$

D'autre part, si E est définie positive on peut écrire

$$(tr\ E)^2 = (\sum_i \lambda_i)^2 \leq \sum_i \lambda_i^2 = tr\ E^2 \tag{II.4}$$

où les λ_i sont les valeurs propres de E. Si l'on pose $E = (S_{12}S_{21})^{1/2}$ dans (II.4) on obtient

$$tr\ S_{12}S_{21} \leq (tr\ (S_{12}S_{21})^{1/2})^2. \tag{II.5}$$

En combinant (II.3) et (II.5) on obtient

$$RV \leq \sqrt{pq} RLS^2. \tag{II.6}$$

Si, d'autre part, l'on pose $E = (S_{12}S_{21})^{1/2}$ et $E = (S_{21}S_{12})^{1/2}$ respectivement dans (II.2) il vient

$$(tr\ (S_{12}S_{21})^{1/2})^2 \leq p\ tr\ S_{12}S_{21} \tag{II.7}$$

et

$$(tr\ (S_{12}S_{21})^{1/2})^2 = (tr\ (S_{21}S_{12})^{1/2})^2 \leq q\ tr\ S_{21}S_{12} = q\ tr\ S_{12}S_{21} \tag{II.8}$$

En posant $E = S_{11}$ et $E = S_{22}$ respectivement dans (II.4) il vient

$$\begin{aligned} (tr\ S_{11})^2 &\geq tr\ S_{11}^2 \\ (tr\ S_{22})^2 &\geq tr\ S_{22}^2 \end{aligned} \tag{II.9}$$

En combinant (II.7) et (II.9) on obtient

$$RLS^2 \leq p\ RV. \tag{II.10}$$

En combinant ensuite (II.8) et (II.9) on obtient également

$$RLS^2 \leq q\ RV. \tag{II.11}$$

Par suite, (II.10) et (II.11) donnent

$$RLS^2 \leq \sqrt{pq}\ RV \tag{II.12}$$

et finalement, (II.6) et (II.12) conduisent au résultat désiré. Notons que, compte tenu de (II.10) et (II.11) on a aussi $RLS^2 \leq \text{Min}\ (p, q)\ RV$, inégalité plus précise.

Bibliographie

[1] CARROLL J.D. Generalization of Canonical Correlation Analysis to Three or More Sets of Variables, Proc. 76 th Annual Convention APA, 1968, 227-228.

[2] ESCOUFIER Y. Le traitement des variables vectorielles, Biometrics, 1973, 29, 751-760.

[3] GREEN B.F. Best Linear Composites with a Specified Structure, Psychometrika, 1969, 34, 301-318.

[4] JOHNSON R.A. and WICHERN D.W. Applied Multivariate Statistical Analysis, Prentice-Hall Inc., New Jersey, 1982.

[5] KETTENRING J.R. Canonical Analysis of Several Sets of Variables, Biometrika, 1971, 58, 433-451.

[6] KIERS, HENK A.L., CLEROUX R. and TEN BERGE, JOS M.F. Generalized Canonical Analysis Based on Optimizing Matrix Correlations and a Relation with IDIOSCAL, submitted for publication, 1991.

[7] LAZRAQ A. et CLEROUX R. Etude comparative de différentes mesures de liaison entre deux vecteurs aléatoires, Stat. et Anal. des Données, 1988A, 13, 15-38.

[8] LAZRAQ A. et CLEROUX R. Un algorithme pas à pas de sélection de variables en régression linéaire multivariée, Stat. et Anal. des Données, 1988B, 13, 39-58.

[9] LINGOES J.C. and SCHÖNEMANN P.H. Alternative Measures of Fit for the Schönemann-Carroll Matrix Fitting Algorithm, Psychometrika, 1974, 39, 423-427.

[10] RAO C.R. Linear Statistical Inference and its Applications, J. Wiley and Sons, New York, 1965.

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

[12] STEWART D. and LOVE W. A General Canonical Correlation Index, Psycho. Bull., 1968, 70, 160-163.