REVUE DE STATISTIQUE APPLIQUÉE

V. NZOBOUNSANA
T. DHORNE

Écart : une nouvelle méthode d'analyse canonique généralisée (ACG)

Revue de statistique appliquée, tome 51, n° 4 (2003), p. 57-82
<http://www.numdam.org/item?id=RSA_2003__51_4_57_0>

Rev. Statistique Appliquée, 2003, **LI** (4), 57-82

ECART : UNE NOUVELLE MÉTHODE D'ANALYSE CANONIQUE GÉNÉRALISÉE (ACG)

V. NZOBOUNSANA[1], T. DHORNE[2]

[1] *Université de RENNES 2*
[2] *Université de Bretagne Sud*

RÉSUMÉ

L'article a pour objet de présenter, d'étudier une nouvelle méthode d'analyse canonique généralisée que nous appelons méthode ECART et la comparer aux méthodes existantes : la méthode MAXVAR et la méthode MINVAR. La méthode consiste en la détermination des variables canoniques, combinaisons linéaires des variables de chaque groupe, telles que la différence entre la plus grande et la plus petite valeur propre de la matrice des corrélations de ces variables canoniques soit maximale. Des propriétés ainsi que les solutions de la méthode Ecart sont données. Une comparaison théorique et numérique avec la méthode MAXVAR proposée par [Horst 61a] et, avec la méthode MINVAR proposée par [Kenttenring 71] est effectuée. Trois jeux de données sont exploités pour illustrer et comparer les 3 méthodes d'ACG en présence.

Mots-clés : *Analyse Canonique, Variables Canoniques, Matrice de corrélation.*

ABSTRACT

This paper presents a new method of generalized canonical correlation analysis (ACG). This method is called method ECART. It consists to maximize the difference between the upper and lower eigenvalues of the correlation matrix of the canonical variates (the linear combination of the variables of each group). The solutions of this method are presented and compared (theoritical and numerical study) with two usual methods : the MAXVAR procedure which is proposed by [Horst 61a] and the MINVAR procedure which is proposed by [Kenttenring 71]. Three data sets are used to compare numerically the three methods of ACG.

Keywords : *Canonical analysis, Canonical variate, Correlation matrix.*

1. Introduction

L'ACG est une technique qui permet d'étudier les liaisons linéaires entre plusieurs groupes de variables mesurées sur les mêmes individus. En général, les critères à optimiser proposés dans la littérature pour atteindre cet objectif sont des fonctions des valeurs propres de la matrice des corrélations des corrélations des variables canoniques (les combinaisons linéaires des variables des groupes). Ces méthodes sont regroupées sous le nom d'Analyse Canonique Généralisée (ACG), car étendant, à un nombre de groupes de variables quelconques, l'analyse canonique proposée par [Hottelling 36].

Dans cet article, nous proposons une nouvelle méthode d'ACG, appelée méthode ECART. Elle consiste à chercher, pour chaque groupe de variables, des variables canoniques de différents ordre et telles que pour un ordre donné la différence entre la plus grande et la plus petite valeur propre de la matrice de corrélations entre variables canoniques (de cet ordre là) soit maximale. Après avoir décrit et donné les solutions de cette nouvelle méthode, nous la comparons avec deux méthodes usuelles d'ACG : la méthode MINVAR, proposée par [Kenttenring 71], qui consiste à minimiser la plus grande valeur propre de la matrice de corrélations canoniques et, la méthode MAXVAR, proposée par [Horst 61a], qui consiste à maximiser la plus petite valeur propre non nulle de la matrice de corrélations canoniques.

Dans la section 2, nous décrivons les méthodes ECART, MAXVAR et MINVAR. Nous montrons que les solutions des méthodes MAXVAR et MINVAR sont obtenues en faisant une décomposition spectrale de la matrice des corrélations des données initiales, tandis que celles de la méthode ECART sont obtenues par une méthode itérative. Elle donne aussi l'algorithme qui permet d'avoir les solutions de la méthode ECART.

Dans la section 3, nous faisons une comparaison des 3 méthodes d'ACG. Cette comparaison est effectuée de deux façons. La première est théorique et la seconde numérique. Trois jeux de données sont exploités pour illustrer et comparer les 3 méthodes d'ACG en présence. Ceci permet d'apprécier la variabilité des résultats de ces trois méthodes.

2. Analyses Canoniques Généralisées

Considérons p groupes de variables quantitatives $X^1, X^2, \cdots, X^p$ mesurées sur les mêmes individus. On suppose que les variables dans les groupes sont centrées, ce qui ne nuit pas à la généralité. Notons m_i le nombre total des variables de X^i, $Z_i = X^i P_i$ une combinaison linéaire des variables de X^i, P_i est un vecteur de R^{m_i}. Notons $Cor(Z_i, Z_j)$ la corrélation linéaire entre Z_i et Z_j et Φ, la matrice $p \times p$ des corrélations linéaires des Z_i. Notons $\lambda_1(\Phi), \lambda_2(\Phi), \cdots, \lambda_p(\Phi)$, les valeurs propres de Φ rangées par ordre décroissant (*i.e.* $\lambda_1(\Phi) \geqslant \lambda_2(\Phi) \geqslant \cdots \geqslant \lambda_p(\Phi)$) et Σ_{ij}, la matrice d'intercorrélation entre les variables de X^i et celles de X^j.

2.1. Méthode ECART

Nous appelons Analyse Canonique Généralisée, selon la méthode ECART, de p groupes de variables X^1, X^2, $\cdots$, X^p, la recherche, pour chaque groupe de variables, des Z_i avec $Z_i = X^i P_i$ et $P_i \in R^{m_i}$ pour $i \in \{1, 2, \cdots, p\}$, normées, appelés variables canoniques telles que la différence entre la plus grande et la plus petite valeur propre de Φ soit maximale. Mathématiquement cela est équivalent à résoudre le problème suivant :

$$(P_1, P_2, \cdots, P_p) = \underset{P_1, P_2, \cdots, P_p}{\arg.\max} \{\frac{1}{p}\{\lambda_1(\Phi) - \lambda_p(\Phi)\}\} \qquad (1)$$

Soit $Z_1^{(1)}, Z_2^{(1)}, \cdots, Z_p^{(1)}$, avec pour tout i, $Z_i^{(1)} = X^i P_i^{(1)}$, l'ensemble des premières variables canoniques solutions de la méthode ECART. Pour compléter cette solution, nous cherchons p nouvelles combinaisons linéaires normées des variables notées $Z_1^{(2)}, Z_2^{(2)}, \cdots, Z_p^{(2)}$, avec pour tout i $Z_i^{(2)} = X^i P_i^{(2)}$ et $P_i^{(2)} \in R^{m_i}$, solutions de (1) et telles que $Z_i^{(2)} = X^i P_i^{(2)}$ soit non corrélée avec $Z_i^{(1)} = X^i P_i^{(1)}$.

Cette procédure peut être répétée au plus $r = min(m_1, m_2, \cdots, m_p)$ fois et donne pour chaque groupe de variables, un ensemble de variables canoniques deux à deux orthogonales.

La proposition suivante donne deux problèmes d'optimisation qui sont équivalents au problème défini par (1).

PROPOSITION 2.1. – *Soit*

- $\Delta_\alpha = Diag(\alpha_1, \alpha_2, \cdots, \alpha_p)$ *la matrice diagonale associée au vecteur* $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_p)$.
- $H_p = \{U : {}^tUU = Id_p\}$, *l'ensemble des matrices orthogonales.*
- $m_{ij}(\alpha)$ *: le terme général de la matrice* $U\Delta_\alpha {}^tU$.
- α^* *le vecteur tel que* $\alpha_1^* = 2$, $\alpha_p^* = 0$ *et pour* $i \in \{2, 3, \cdots, p-1\}$, $\alpha_i^* = 1$.

Les problèmes (2) et (3), de détermination de $(P_1, P_2, \cdots, P_p)$, *sont équivalents au problème (1) :*

$$(P_1, P_2, \cdots, P_p, U) = \underset{P_1, P_2, \cdots, P_p}{\arg.\max} \max_{U \in H_p} \{\frac{1}{p} tr[{}^tU\Phi U\Delta_{\alpha^*}]\} \tag{2}$$

$$(P_1, P_2, \cdots, P_p, U) = \underset{P_1, P_2, \cdots, P_p}{\arg.\max} \max_{U \in H_p} \{\frac{1}{p} \sum_{i=1}^{p} \sum_{j=1}^{p} m_{ij}(\alpha^*) Cor(Z_i, Z_j)\}. \tag{3}$$

Le lemme suivant permet de démontrer les résultats de la proposition 2.1.

LEMME 2.1 [Anderson 84]. – *Soit* Φ *une matrice* $p \times p$ *symétrique et définie positive. Soit* Δ_α *une matrice diagonale de dimension* $p \times p$ *dont le* i*-ième élément de la diagonale principale est* α_i *(positif), alors on a le résultat suivant :*

$$\max_{U \in H_p} [\frac{1}{p} tr\{{}^tU\Phi U\Delta_\alpha\}] = \frac{1}{p} \sum_{i=1}^{p} \alpha_i \lambda_i(\Phi), \tag{4}$$

où H_p *désigne l'ensemble des matrices orthogonales d'ordre* p. *Le maximum de (4) est atteint lorsque* U *est la matrice, de dimension* $p \times p$, *ayant pour colonnes les vecteurs propres de* Φ.

DÉMONSTRATION. – Les assertions de la proposition 2.1 se démontrent aisément.

- Premièrement, l'équivalence entre (1) et (2) est directe. En effet, Φ est une matrice de corrélation d'ordre p et donc la somme de ses p valeurs propres est égale à p. Cela entraîne donc,

$$\lambda_1(\Phi) - \lambda_p(\Phi) = \{2\lambda_1(\Phi) + \lambda_2(\Phi) + \cdots + \lambda_{p-1}(\Phi)\} - p \tag{5}$$

et par conséquent, maximiser $\{\frac{1}{p}[\lambda_1(\Phi) - \lambda_p(\Phi)]\}$ par rapport à $P_1, P_2, \cdots, P_p$ est équivalent à maximiser par rapport aux mêmes vecteurs la fonction $\frac{1}{p}[2\lambda_1(\Phi) + \lambda_2(\Phi) + \cdots + \lambda_{p-1}(\Phi)]$.

En appliquant les résultats du lemme 2.1 à la fonction précédente on déduit que

$$\max_{P_1,P_2,\cdots,P_p} \{\frac{1}{p}[\lambda_1(\Phi) - \lambda_p(\Phi)]\} \iff \max_{P_1,P_2,\cdots,P_p,U} \{\frac{1}{p} tr[{}^tU\Phi U\Delta_\alpha]\} \quad (6)$$

où ${}^tUU = Id_p$ et $\Delta_\alpha = diag(2, 1, 1, \cdots, 1, 0)$.

• Deuxièmement, l'équivalence entre (2) et (3) s'obtient en utilisant les propriétés de la fonction trace et celles du produit des matrices.

2.1.1. Solutions de la méthode ECART

Dans cette sous-section, nous donnons les solutions de la méthode ECART. Nous supposons que la matrice U est fixée et que, sans perte de généralité, $\Sigma_{ii} = Id_{m_i}$ (quitte à raisonner, par un changement des variables, sur les $Y_i = X^i\Sigma_{ii}^{-\frac{1}{2}}$).

Solution analytique d'ordre un

La proposition suivante donne les solutions d'ordre un de la méthode ECART.

PROPOSITION 2.2. – *Les solutions de la méthode ECART sous la contrainte que $Z_1, Z_2, \cdots, Z_p$ sont normées, sont données par les égalités suivantes :*

$$\lambda_i^* P_i = \frac{1}{p} \sum_{j=1, j\neq i}^{p} m_{ij}(\alpha)\Sigma_{ij}P_j \quad \text{pour tout } i \in \{1, 2, \cdots, p\} \quad (7)$$

où $\lambda_1^*, \lambda_2^*, \cdots, \lambda_p^*$ sont des multiplicateurs de Lagrange.

Cette proposition définit une méthode itérative d'obtention des solutions de la méthode ECART.

DÉMONSTRATION. – Nous avons vu dans la proposition 2.1 que les solutions de la méthode ECART sont obtenues en cherchant simultanément les vecteurs canoniques $P_1, P_2, \cdots, P_p$ et une matrice orthogonale U tels que :

$$\frac{1}{p} tr[{}^tU\Phi U\Delta_\alpha] = \frac{1}{p} \sum_{i=1}^{p}\sum_{j=1}^{p} m_{ij}(\alpha)Cor(Z_i, Z_j) \quad (8)$$

soit maximale sous les contraintes $Var(Z_i) = {}^tP_i\Sigma_{ii}P_i = {}^tP_iP_i = 1$ pour tout i.

En fixant la matrice U et en introduisant les multiplicateurs de Lagrange λ_1, $\lambda_2, \cdots, \lambda_p$, on doit donc maximiser la quantité :

$$L(P_1, P_2, \cdots, P_p, U) = \frac{1}{p} \sum_{i=1}^{p} \sum_{j=1}^{p} m_{ij}(\alpha) Cor(Z_i, Z_j) - \sum_{i=1}^{p} \lambda_i (Var(Z_i) - 1). \tag{9}$$

En dérivant L par rapport à chaque P_i et en égalant le résultat à zéro, on déduit les égalités (7) :

$$\lambda_i^* P_i = \frac{1}{p} \sum_{j=1, j \neq i}^{p} m_{ij}(\alpha) \Sigma_{ij} P_j \quad où \quad \lambda_i^* = \lambda_i - \frac{m_{ii}(\alpha)}{p}. \tag{10}$$

Par ailleurs, nous déduisons de la contrainte de normalisation des variables canoniques et des égalités (10) que

$$\lambda_i = \frac{m_{ii}(\alpha)}{p} + \frac{1}{p} \sum_{j=1, j \neq i}^{p} m_{ij}(\alpha)^t P_i \Sigma_{ij} P_j = \frac{1}{p} \sum_{j=1}^{p} m_{ij}(\alpha) Cor(Z_i, Z_j), \tag{11}$$

d'où la proposition. □

Solution analytique d'ordre supérieur

Après avoir montré comment trouver les solutions d'ordre un du problème défini par (1), nous donnons, ici, les solutions d'ordre s ($s > 1$) problème (1).

DÉFINITION 2.1. – *Nous appelons par solutions d'ordre k ($k > 1$) du problème (1), la recherche de p combinaisons linéaires $Z_1^{(k)}, Z_2^{(k)}, \cdots, Z_p^{(k)}$ où pour tout i, $Z_i^{(k)} = X^i P_i^{(k)}$, $P_i^{(k)} \in R^{m_i}$ et $\Phi = (Cor(Z_i^{(k)}, Z_j^{(k)}))_{ij}$ telles que*

$$(P_1^{(k)}, P_2^{(k)}, \cdots, P_p^{(k)}) = arg. \max_{P_1^{(k)}, P_2^{(k)}, \cdots, P_p^{(k)}} \max_{U} \{tr[{}^t U \Phi U \Delta_\alpha]\}, \tag{12}$$

sous les contraintes :

$$Var(Z_1^{(k)}) = Var(Z_2^{(k)}) = \cdots = Var(Z_p^{(k)}) = 1 \quad et \tag{13}$$

$$Cor(Z_i^{(k)}, Z_i^{(s)}) = 0 \quad pour \quad s \in \{1, 2, \cdots, (k-1)\}, \quad k \in \{2, \cdots, r\}, \tag{14}$$

où Δ_α est une matrice diagonale définit par $\Delta_\alpha = diag(2, 1, 1, \cdots, 1, 0)$.

Dans la proposition suivante, nous donnons les égalités que doivent vérifier les solutions d'ordre supérieur de l'ACG selon la méthode ECART.

PROPOSITION 2.3. – *Soit $Z_i^{(k)}$ les solutions d'ordre k du problème (1). Sous la contrainte de normalisation (13) et les contraintes additionnelles (14), les solutions du problème (1) à l'ordre s ($s > 1$), $k \in \{1, 2, 3, \cdots, (s-1)\}$, vérifient les p équations suivantes :*

$$\lambda_i^* P_i^{(s)} = \sum_{j=1, j\neq i}^{p} m_{ij}(\alpha)[Id_{m_i} - \sum_{k=1}^{s-1} P_i^{(k)t} P_i^{(k)}]\Sigma_{ij}[Id_{m_j} - \sum_{k=1}^{s-1} P_j^{(k)t} P_j^{(k)}]P_j^{(s)} \quad \forall i, \tag{15}$$

où $\lambda_1^*, \lambda_2^*, \cdots, \lambda_p^*$ sont des constantes de normalisation (des multiplicateurs de Lagrange) et $m_{ij}(\alpha)$, pour tout i et j, est le terme général de la matrice $M = U\Delta_\alpha {}^tU$.

DÉMONSTRATION. – Comme pour le calcul des solutions d'ordre un, a l'ordre s ($s > 1$), les solutions de la méthode ECART sont obtenues en maximisant la fonction suivante :

$$L(P_1^{(s)}, P_2^{(s)}, \cdots, P_p^{(s)}) = \frac{1}{p}\sum_{i=1}^{p}\sum_{j=1}^{p} m_{ij}(\alpha) Cor(Z_i^{(s)}, Z_j^{(s)}) - \sum_{i=1}^{p} \lambda_i (Var(Z_i^{(s)}) - 1) - 2\sum_{i=1}^{p}\sum_{k=1}^{s-1} \beta_i^k Cor(Z_i^{(k)}, Z_i^{(s)}), \tag{16}$$

où $\lambda_1, \lambda_2, \cdots, \lambda_p$ (respectivement les $\beta_1^k, \beta_2^k, \cdots, \beta_p^k$ pour tout $k \in \{1, 2, \cdots, (s-1)\}$) sont des multiplicateurs de Lagrange.

Il est facile de montrer en utilisant les contraintes $Var(Z_i^l) = {}^tP_i^{(l)} P_i^{(l)} = 1$ pour tout i et pour tout $l \in \{1, 2, \cdots, (s-1)\}$, que la fonction L est égale à :

$$L(P^{(s)}) = \sum_{i=1, j=1, i\neq j}^{p} \frac{m_{ij}}{p}(\alpha) {}^tP_i^{(s)} \Sigma_{ij} P_j^{(s)} - \sum_{i=1}^{p} \lambda_i^{*t} P_i^{(s)} P_i^{(s)} + \sum_{i=1}^{p} \lambda_i - 2\sum_{i=1}^{p}\sum_{k=1}^{s-1} \beta_i^{kt} P_i^{(k)} P_i^{(s)}, \tag{17}$$

où $P^{(s)} = (P_1^{(s)}, P_2^{(s)}, \cdots, P_p^{(s)})$ et $\lambda_i^* = \{\lambda_i - \frac{m_{ii}(\alpha)}{p}\}$.

En dérivant L par rapport à chaque $P_i^{(s)}$ et en égalant le résultat à 0, on obtient :

$$\frac{1}{p}\sum_{j=1, j\neq i}^{p} m_{ij}(\alpha)\Sigma_{ij} P_j^{(s)} - \lambda_i^* P_i^{(s)} - \sum_{k=1}^{s-1} \beta_i^k P_i^{(k)} = 0, \tag{18}$$

En multipliant successivement, $\forall k$, l'équation (18) à gauche, par ${}^tP_i^{(k)}$, on obtient

$$\beta_i^k = \frac{1}{p} \sum_{j=1, j\neq i}^{p} m_{ij}(\alpha) {}^tP_i^{(k)} \Sigma_{ij} P_j^{(s)} \quad \forall k \in \{1, 2, \cdots, (s-1)\}, \tag{19}$$

En remplaçant les β_i^k dans l'équation (18), on a :

$$\frac{1}{p} \sum_{j=1, j\neq i}^{p} m_{ij}(\alpha) \Sigma_{ij} P_j^{(s)} - \lambda_i^* P_i^{(s)} - \frac{1}{p} \sum_{j=1, j\neq i}^{p} m_{ij}(\alpha) [\sum_{k=1}^{s-1} P_i^{(k)} {}^tP_i^{(k)}] \Sigma_{ij} P_j^{(s)} = 0, \tag{20}$$

ou encore en réordonnant les trois termes de (20) :

$$\frac{1}{p} \sum_{j=1, j\neq i}^{p} m_{ij}(\alpha) [Id_{m_i} - \sum_{k=1}^{s-1} P_i^{(k)} {}^tP_i^{(k)}] \Sigma_{ij} P_j^{(s)} - \lambda_i^* P_i^{(s)} = 0, \tag{21}$$

donc $\forall i$, les $P_i^{(s)}$, pour U et Δ_α fixées, vérifient les p égalités suivantes :

$$\lambda_i^* P_i^{(s)} = \frac{1}{p} \sum_{j=1, j\neq i}^{p} m_{ij}(\alpha) [Id_{m_i} - \sum_{k=1}^{s-1} P_i^{(k)} {}^tP_i^{(k)}] \Sigma_{ij} P_j^{(s)} \quad \forall i \in \{1, 2, \cdots, p\}, \tag{22}$$

où de manière équivalente :

$$\lambda_i^* P_i^{(s)} = \frac{1}{p} \sum_{j=1, j\neq i}^{p} m_{ij}(\alpha) [Id_{m_i} - \sum_{k=1}^{s-1} P_i^{(k)} {}^tP_i^{(k)}] \Sigma_{ij} [Id_{m_j} - \sum_{k=1}^{s-1} P_j^{(k)} {}^tP_j^{(k)}] P_j^{(s)} \tag{23}$$

$\forall i \in \{1, 2, \cdots, p\}$ puisque $[Id_{m_i} - \sum_{k=1}^{s-1} P_i^{(k)} {}^tP_i^{(k)}] P_i^{(s)} = P_i^{(s)}$. D'où la proposition. □

2.1.2. Algorithme numérique

L'algorithme présenté ici et appelé dans la suite algorithme S est déduit des égalités (10) et (23) (puisque ces deux équations montrent que les P_i sont solutions d'un système d'équations non linéaires). Il s'agit, pour $\Delta_\alpha = diag(2, 1, 1, 1, \cdots, 1, 0)$, de

1. Choisir p vecteurs canoniques P_i de départ où $P_i \in R^{m_i}$ pour tout i,
2. Pour $P_1, P_2, \cdots, P_p$ fixés, calculer les vecteurs et valeurs propres de la matrice de corrélation Φ, $\Phi = V \delta_\lambda {}^tV$, $\delta_\lambda = diag(\lambda_1(\Phi), \cdots, \lambda_p(\Phi))$ et prendre $U = V$.

3. Calculer les vecteurs canoniques itérativement selon le système ci-dessous :

$$S = \begin{cases} P_1' = \frac{1}{p}\sum_{j,j\neq 1}^{p} m_{1j}(\alpha)\Sigma_{1j}P_j, \quad \lambda_1 = ||P_1'||, \quad P_1 = \frac{P_1'}{\lambda_1} \\ P_2' = \frac{1}{p}\sum_{j=1,j\neq 2}^{p} m_{2j}(\alpha)\Sigma_{2j}P_j, \quad \lambda_2 = ||P_2'||, \quad P_2 = \frac{P_2'}{\lambda_2} \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ P_{p-1}' = \frac{1}{p}\sum_{j=1,j\neq(p-1)}^{p} m_{(p-1)j}(\alpha)\Sigma_{(p-1)j}P_j, \quad \lambda_{p-1} = ||P_{p-1}'||, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad P_{p-1} = \frac{P_{p-1}'}{\lambda_{p-1}} \\ P_p' = \frac{1}{p}\sum_{j=1,j\neq p}^{p} m_{pj}(\alpha)\Sigma_{pj}P_j, \quad \lambda_p = ||P_p'||, \quad P_p = \frac{P_p'}{\lambda_p} \end{cases}$$

4. Avec arrêt lorsque pour ϵ positif fixé et petit, on a l'inégalité suivante

$$\sum_{i=1}^{p} |\lambda_i^{(n)} - \lambda_i^{(n+1)}| < \epsilon.$$

Pour obtenir numériquement les solutions d'ordre s ($s > 1$) de la méthode ECART, on utilise le même algorithme S. On remplace, pour tout i et j, les matrices Σ_{ij} par des nouvelles matrices $[Id_{m_i} - \sum_{k=1}^{s-1} P_i^{(k)t} P_i^{(k)}]\Sigma_{ij}[Id_{m_j} - \sum_{k=1}^{s-1} P_j^{(k)t} P_j^{(k)}]$.

2.2. *Méthodes MINVAR et MAXVAR*

Les méthodes MINVAR [Kenttenring 71] et MAXVAR [Horst 61a] consistent dans un premier temps, à trouver p combinaisons linéaires normées $Z_1, Z_2, \cdots, Z_p$ (solutions d'ordre un) des variables de chaque groupe telles que la plus petite (respectivement la plus grande) valeur propre de la matrice de corrélation canonique soit minimale (respectivement maximale).

Les solutions d'ordre s ($s > 1$)) consistent à trouver également p combinaisons linéaires normées des variables de chaque groupe telles que les mêmes critères soient optimisés, sous les contraintes additionnelles que, les variables canoniques d'ordre s soient orthogonales aux variables canoniques d'ordre $(s-1)$ pour $s \in \{2, 3, \cdots, r\}$ et $r = min(m_1, m_2, \cdots, m_p)$.

DÉFINITION 2.2. – *Soit Φ une matrice $p \times p$ symétrique et définie positive. Soient $\lambda_1(\Phi) \geqslant \lambda_2(\Phi) \geqslant \cdots \geqslant \lambda_p(\Phi) > 0$ ses p valeurs propres non nulles où $\lambda_1(\Phi)$ (respectivement $\lambda_p(\Phi)$) définit sa plus grande (respectivement plus petite) valeur propre. On a les résultats suivants (cf [Anderson 84])*

$$\lambda_1(\Phi) = \max_w \{{}^t w\Phi w\} \quad et \quad \lambda_p(\Phi) = \min_u \{{}^t u\Phi u\}, \tag{24}$$

où w (respectivement u) est un vecteur normé défini par $w = {}^t\{w_1, w_2, \cdots, w_p\} \in R^p$ (respectivement $u = {}^t\{u_1, u_2, \cdots, u_p\} \in R^p$).

PROPOSITION 2.4 [Kenttenring 71]. – *Sous la contrainte* $Var(Z_i) = 1$ *pour tout* i*, les vecteurs* P_i*, appelés vecteurs canoniques, sont solutions de la méthode MAXVAR ou MINVAR si et seulement si, ils vérifient les égalités suivantes :*

$$P_i^{(1)} = \frac{w_{max}^i}{({}^t w_{max}^i w_{max}^i)^{\frac{1}{2}}} \quad \textit{pour MAXVAR et}$$

$$P_i^{(1)} = \frac{u_{min}^i}{({}^t u_{min}^i u_{min}^i)^{\frac{1}{2}}} \quad \textit{pour MINVAR} \tag{25}$$

avec

- u_{min} *(respectivement* w_{max}*) le dernier (respectivement le premier) vecteur propre de* Σ*.*
- u_{min}^i *et* w_{max}^i *les i-ièmes sous vecteurs de* u_{min} *et de* w_{max} *de longueur* m_i*.*
- Σ *la matrice des corrélations des groupes de variables* $Y^1, Y^2, \cdots, Y^p$ *avec pour tout* i*,* $Y^i = X^i \Sigma_{ii}^{-1/2}$ *et* Σ_{ii}*, la matrice de corrélation des variables de* X^i*.*

Les variables canoniques de la méthode MAXVAR (respectivement de la méthode MINVAR) vérifient les égalités ci-dessous, pour $i \in \{1, 2, \cdots, p\}$ *:*

$$Z_i^{(1)} = \frac{X^i \Sigma_{ii}^{-\frac{1}{2}} w_{max}^i}{({}^t w_{max}^i w_{max}^i)^{\frac{1}{2}}} \quad \textit{pour MAXVAR} \tag{26}$$

$$Z_i^{(1)} = \frac{X^i \Sigma_{ii}^{-\frac{1}{2}} u_{min}^i}{({}^t u_{min}^i u_{min}^i)^{\frac{1}{2}}} \quad \textit{pour MINVAR} \tag{27}$$

Enfin, le maximum de la méthode MAXVAR est égal à la plus grande valeur propre de Σ*, tandis que le minimum de la méthode MINVAR est égal à la dernière valeur propre non nulle de* Σ*.*

Soit $P_i^{(h)}$ les solutions d'ordre h de la méthode MAXVAR (respectivement de la méthode MINVAR), la proposition suivante donne les solutions d'ordre s des deux méthodes.

PROPOSITION 2.5 [Kenttenring 71]. – *On impose la contrainte que les variables canoniques des groupes à l'ordre* s *sont orthogonales aux variables canoniques trouvées à l'ordre* h $(h \in \{1, 2, \cdots, (s-1)\})$*. A l'ordre* s*, les solutions de la méthode MAXVAR (respectivement de la méthode MINVAR) vérifient les égalités de la proposition 2.4 avec* Σ *remplacée par une nouvelle matrice* $M = M^{(s)} \Sigma M^{(s)}$ *où* $M^{(s)}$ *est égale à la matrice bloc diagonale ayant pour terme général la matrice* M_{ii} *avec, pour tout* $i \in \{1, 2, \cdots, p\}$*,* $M_{ii} = [Id_{m_i} - \sum_{h=1}^{(s-1)} P_i^{(h)t} P_i^{(h)}]$*.*

• **Algorithme**

Soit $P_{1max}^{(h)}, P_{2max}^{(h)}, \cdots, P_{pmax}^{(h)}$ et $P_{1min}^{(h)}, P_{2min}^{(h)}, \cdots, P_{pmin}^{(h)}$ les solutions d'ordre h où $h \in \{1, 2, \cdots, s\}$ de la méthode MAXVAR et celles de la méthode MINVAR. Les solutions d'ordre $(s+1)$ de ces 2 méthodes peuvent être obtenues selon l'algorithme ci-dessous :

Première étape : Construction des projecteurs $\Pi_k^{(1)}$ et $\Pi_k^{(p)}$ sur les sous-espaces vectoriels engendrés par les s premiers vecteurs canoniques de X^k obtenus avec la méthode MAXVAR (respectivement avec la méthode MINVAR).

Deuxième étape : Définir p nouveaux groupes dont la matrice globale est :

• $X_1^{(2)}$ (pour calculer les solutions d'ordre $(s+1)$ de la méthode MAXVAR) où

$$X_1^{(2)} = [Y^1(Id - \Pi_1^{(1)})|Y^2(Id - \Pi_2^{(1)})|\cdots|Y^k(Id - \Pi_k^{(1)})|\cdots|Y^p(Id - \Pi_p^{(1)})] \tag{28}$$

• $X_p^{(2)}$ (pour calculer les solutions d'ordre $(s+1)$ de la méthode MINVAR) où

$$X_p^{(2)} = [Y^1(Id - \Pi_1^{(p)})|Y^2(Id - \Pi_2^{(p)})|\cdots|Y^k(Id - \Pi_k^{(p)})|\cdots|Y^p(Id - \Pi_p^{(p)})] \tag{29}$$

Troisième étape : Calculer les solutions d'ordre 1 en remplaçant Σ (la matrice des corrélations des Y^i) par $\Sigma_{x_1^{(2)}}$ (respectivement par $\Sigma_{x_p^{(2)}}$) pour les solutions d'ordre $(s+1)$ de la méthode MAXVAR (respectivement pour les solutions d'ordre $(s+1)$ de la méthode MINVAR) où $\Sigma_{x_1^{(2)}}$ (respect. par $\Sigma_{x_p^{(2)}}$) est la matrice des corrélations des variables de $X_1^{(2)}$ (respectivement de $X_p^{(2)}$).

3. Comparaison des 3 méthodes d'ACG

3.1. Comparaison théorique

• Les trois méthodes d'ACG, présentées ici, ont un point commun. Elles déterminent, pour chaque groupe de variables, un ensemble de variables canoniques normées et deux à deux orthogonales (*cf* les contraintes 13 et 14). Ceci permet de comparer la dispersion des individus selon chaque groupe de variables.

Elles donnent souvent un éclairage différent. En effet, pour la méthode MAXVAR (respectivement MINVAR) les solutions sont obtenues en maximisant la plus grande (respectivement en minimisant la dernière ou encore en maximisant la somme des $p-1$ premières) valeur(s) propre(s) de Φ tandis que, les solutions la méthode ECART sont obtenues en maximisant la différence entre la plus grande et la plus petite valeur propre de Φ. De plus, la valeur optimale de la méthode MAXVAR est égale à la plus grande valeur propre de Σ tandis que, celle de la méthode MINVAR est égale à la dernière valeur propre non nulle de cette même matrice.

• (i) Lorsque la dernière valeur propre non nulle de Σ est multiple, la méthode MINVAR ne donne pas une solution unique, c'est-à-dire que, les vecteurs canoniques optimaux ne sont pas uniques.

(ii) De même, lorsque la plus grande valeur propre de Σ est multiple, les solutions de la méthode MAXVAR ne sont pas stables.

En effet, si la dernière valeur propre non nulle (respectivement la plus grande valeur propre) de Σ est multiple, Σ admet plusieurs vecteurs propres associés à la dernière valeur propre non nulle (respectivement à la plus grande valeur propre). Comme cette dernière n'est pas unique et que les solutions de la méthode MINVAR (respectivement de la méthode MAXVAR) sont des fonctions de ces vecteurs propres alors, ces dernières ne sont pas aussi uniques.

• Lorsque la première valeur propre de Σ est égale au nombre de groupes de variables, la méthode ECART et la méthode MAXVAR donnent les mêmes solutions d'ordre un.

En effet, lorsque la première valeur propre de Σ est égale à p (p, le nombre de groupe), Φ admet une plus grande valeur propre égale à p et les autres valeurs propres sont nulles. Cela entraîne que, maximiser la plus grande valeur propre de Φ revient à maximiser la différence entre la plus grande et la plus petite valeur propre de Φ. D'où le résultat.

Remarque 3.1. – Les solutions des 3 méthodes d'ACG peuvent être obtenues selon le même algorithme S. Il suffit d'utiliser (4) en prenant $\alpha_1 = 1$ et $\alpha_i = 0$, $i \in \{2, \cdots, p\}$ pour la méthode MAXVAR; $\alpha_i = 1, i \in \{1, 2, \cdots, p-1\}$, et $\alpha_p = 0$ pour la méthode MINVAR et enfin, $\alpha_1 = 2$, $\alpha_i = 1$, $i \in \{2, \cdots, p-1\}$, et $\alpha_p = 0$ pour ECART.

PROPOSITION 3.1. – *Les 3 méthodes d'ACG donnent dans le cas de deux groupes de variables les mêmes solutions que l'analyse canonique linéaire proposée par [Hotelling 36].*

En effet, lorsque le nombre de groupes de variables est égale à deux ($p = 2$), la matrice Φ est de dimension 2×2 et admet, deux valeurs propres $\lambda_1(\Phi)$ (la plus grande) et $\lambda_2(\Phi)$ (la plus petite) avec $\lambda_1(\Phi) = 1 + Cor(Z_1, Z_2)$ et $\lambda_2(\Phi) = 1 - Cor(Z_1, Z_2)$. Par conséquent,

$$\max_{P_1,P_2}\{\lambda_1(\Phi)\} = \max_{P_1,P_2}\{1 + Cor(Z_1, Z_2)\} \Longleftrightarrow \max_{P_1,P_2}\{Cor(Z_1, Z_2)\} \qquad (30)$$

$$\min_{P_1,P_2}\{\lambda_2(\Phi)\} = \min_{P_1,P_2}\{1 - Cor(Z_1, Z_2)\} \Longleftrightarrow \max_{P_1,P_2}\{Cor(Z_1, Z_2)\} \qquad (31)$$

$$\max_{P_1,P_2}\{\frac{1}{2}\{\lambda_1(\Phi) - \lambda_2(\Phi)\}\} = \max_{P_1,P_2}\{Cor(Z_1, Z_2)\} \qquad (32)$$

PROPPRIÉTÉ 3.2. – *La fonction maximisée dans la méthode ECART est une mesure d'association généralisée. Elle est nulle si et seulement si Φ est égale à l'identité (i.e. $Cor(Z_i, Z_j) = 0 \quad \forall i \neq j$), et est égale à 1 si et seulement si toutes*

les valeurs propres de Φ sont nulles exceptée la première qui est égale à p tandis que le critère maximisé dans la méthode MAXVAR n'est jamais nul par définition.

On a en effet,

$$\frac{1}{p}\{\lambda_1(\Phi)-\lambda_p(\Phi)\}=0 \Longleftrightarrow \lambda_1(\Phi)=\lambda_p(\Phi) \Longleftrightarrow \lambda_1(\Phi)=\lambda_2(\Phi)=\cdots=\lambda_p(\Phi) \tag{33}$$

De (33), on en déduit que le critère maximisé dans la méthode ECART est égal à 0 si et seulement si toutes les valeurs propres de Φ sont égales à 1 car $\sum_{i=1}^{p}\lambda_i(\Phi)=p$, $\lambda_1(\Phi)=max(\lambda_1(\Phi),\cdots,\lambda_p(\Phi))$ et $\lambda_p(\Phi)=min(\lambda_1(\Phi),\cdots,\lambda_p(\Phi))$.

$$\frac{1}{p}(\lambda_1(\Phi)-\lambda_p(\Phi))=1 \Longleftrightarrow \lambda_1(\Phi)=p+\lambda_p(\Phi) \Longleftrightarrow \quad \lambda_1(\Phi)=\sum_{i=1}^{p}\lambda_i(\Phi)+\lambda_p(\Phi)$$

$$\Longleftrightarrow \lambda_2(\Phi)+\lambda_3(\Phi)+\cdots+2\lambda_p(\Phi)=0 \tag{34}$$

$$\Longleftrightarrow \lambda_2(\Phi)=\lambda_3(\Phi)=\cdots=\lambda_p(\Phi)=0 \tag{35}$$

car les valeurs propres de Φ sont positives ou nulles et donc (34) est nulle si et seulement si (35) est vraie. De (35), on en déduit que le critère maximisé dans la méthode ECART est égal à 1 si et seulement si la plus grande valeur propre de Φ est égale à p et les autres égales à 0.

3.2. Comparaison numérique

Ici, nous comparons les 3 méthodes d'ACG sur 3 jeux de données. Les 2 premiers exemples utilisent des données proposées dans l'article de Foucart [Foucart 96]. Ces données sont des données simulées (*cf.* [Foucart 96]) tandis que, le troisième exemple concerne des données réelles. Il s'agit des données, de vins, proposées dans l'article de [Chessel 98].

3.2.1. Premier exemple

Pour ce premier exemple, les tableaux 1 et 2 donnent respectivement la matrice des corrélations des variables initiales, Σ, et les valeurs propres de la matrice Σ. Pour ces données (tableau 1), les corrélations linéaires entre les variables de groupes différents sont faibles tandis que, les corrélations entre variables d'un même groupe sont moyennement élevées. La plus grande corrélation canonique entre le premier et le deuxième groupe de variables est égale à 0.172. Elle est égale à 0.167 pour le premier et le troisième groupe et enfin, elle est égale à 0.164 pour le deuxième et troisième groupe.

TABLEAU 1
Matrice de corrélation entre les variables initiales (Σ_x)

	X_1^1	X_2^1	X_3^1	X_1^2	X_2^2	X_3^2	X_1^3	X_2^3	X_3^3
X_1^1	1.000								
X_2^1	0.630	1.000							
X_3^1	0.514	-0.339	1.000						
X_1^2	0.091	0.035	0.071	1.000					
X_2^2	0.114	0.078	0.053	0.583	1.000				
X_3^2	-0.022	-0.044	0.020	0.439	-0.472	1.000			
X_1^3	-0.003	-0.024	0.024	-0.004	0.006	-0.015	1.000		
X_2^3	0.080	0.094	-0.009	0.020	-0.053	0.078	0.591	1.000	
X_3^3	-0.094	-0.130	0.034	-0.022	0.065	-0.098	0.461	-0.441	1.000

TABLEAU 2
Les valeurs propres de la matrice Σ

$\lambda_1(\Sigma)$	$\lambda_2(\Sigma)$	$\lambda_3(\Sigma)$	$\lambda_4(\Sigma)$	$\lambda_5(\Sigma)$	$\lambda_6(\Sigma)$	$\lambda_7(\Sigma)$	$\lambda_8(\Sigma)$	$\lambda_9(\Sigma)$
1.209	1.144	1.109	1.052	1.024	0.955	0.898	0.832	0.773

TABLEAU 3
Vecteurs canoniques d'ordre un

Méthodes	Groupe 1			Groupe 2			Groupe 3		
	X_1^1	X_2^1	X_3^1	X_1^2	X_2^2	X_3^2	X_1^3	X_3^1	X_3^3
MAXVAR	0.950	0.241	-0.194	-0.248	0.732	0.634	0.082	-0.005	-0.996
MINVAR	-0.749	-0.655	0.095	0.091	-0.920	0.376	0.074	-0.677	0.731
ECART	0.915	0.392	-0.093	0.073	0.980	0.182	0.154	0.279	-0.947

Les tableaux 3 et 5 donnent les vecteurs canoniques et les valeurs propres de Φ, des trois groupes de variables, obtenus avec les méthodes MAXVAR, MINVAR et ECART.

Nous constatons, au vu des résultats de ce tableau (tableau 3), que les 3 méthodes d'ACG donnent des vecteurs canoniques qui sont complètement différents. Les valeurs propres de Φ ainsi que, la différence entre la plus grande et la petite valeur propre de Φ (*cf.* tableau 5) sont aussi très différentes d'une méthode à l'autre.

TABLEAU 4
Corrélations entre les variables canoniques d'ordre 1

Méthodes	MAXVAR			MINVAR			ECART		
	Gr[1]	Gr[2]	Gr[3]	Gr[1]	Gr[2]	Gr[3]	Gr[1]	Gr[2]	Gr[3]
Gr[1]	1.000	-0.101	0.153	1.000	0.111	0.137	1.000	0.128	0.151
Gr[2]	-0.101	1.000	-0.049	0.111	1.000	-0.088	0.128	1.000	-0.016
Gr[3]	0.153	-0.049	1.000	0.137	-0.088	1.000	0.151	-0.016	1.000

TABLEAU 5
Les valeurs propres de Φ

Méthodes	$\lambda_1(\Phi)$	$\lambda_2(\Phi)$	$\lambda_3(\Phi)$	$\lambda_1(\Phi) - \lambda_3(\Phi)$
MAXVAR	1.209	0.955	0.836	0.373
ECART	1.191	1.015	0.792	0.398
MINVAR	1.141	1.084	0.773	0.368

La valeur optimale du critère maximisé de la méthode MAXVAR est égale à 1.209 (la plus grande valeur propre de Σ), celle de la méthode MINVAR est égale à 0.773 (la plus petite valeur propre non nulle de Σ) et enfin, celle de la méthode ECART est égale à 0.398.

Dans le tableau 4 nous avons donné, pour chaque méthode, la matrice de corrélations, Φ, obtenues à partir des variables canoniques optimales. En analysant les résultats de ce tableau, nous remarquons que les corrélations linéaires entre les variables canoniques sont très faibles quel que soit la méthode utilisée. Elles ne dépassent pas 0.2 en valeur absolue. Donc, il n'existe pas de liaison linéaire entre les 3 groupes de variables.

Conclusion : Les résultats obtenus avec les 3 méthodes sont cohérents avec les résultats attendus (car les corrélations linéaires entre les variables issues de deux groupes différents étaient très faibles de même que les corrélations canoniques maximales dans les analyses canoniques usuelles de groupes pris 2 à 2). Les 3 méthodes donnent la même interprétation en ce qui concerne la mise en évidence d'une liaison linéaire existante entre les trois groupes de variables (il n'existe pas de liaison linéaire entre les 3 groupes de variables).

3.2.2. Deuxième exemple

Pour ce deuxième exemple, les tableaux 6 et 7 donnent la matrice des corrélations, Σ_x, des variables initiales et les valeurs propres de Σ.

TABLEAU 6
Matrice de corrélations entre les variables initiales (Σ_x)

	X_1^1	X_2^1	X_3^1	X_1^2	X_2^2	X_3^2	X_1^3	X_2^3	X_3^3
X_1^1	1.000								
X_2^1	0.020	1.000							
X_3^1	0.035	0.094	1.000						
X_1^2	0.091	0.080	0.630	1.000					
X_2^2	-0.004	0.591	-0.024	-0.003	1.000				
X_3^2	0.439	0.078	-0.044	-0.022	-0.015	1.000			
X_1^3	0.583	-0.053	0.078	0.114	0.006	-0.472	1.000		
X_2^3	-0.022	-0.441	-0.130	-0.094	0.461	-0.098	0.065	1.000	
X_3^3	0.071	-0.009	-0.339	0.514	0.024	0.020	0.053	0.034	1.000

TABLEAU 7
Les valeurs propres de la matrice Σ

$\lambda_1(\Sigma)$	$\lambda_2(\Sigma)$	$\lambda_3(\Sigma)$	$\lambda_4(\Sigma)$	$\lambda_5(\Sigma)$	$\lambda_6(\Sigma)$	$\lambda_7(\Sigma)$	$\lambda_8(\Sigma)$	$\lambda_9(\Sigma)$
1.8166	1.6402	1.5185	1.4672	1.3531	1.1993	0.0020	0.0015	0.0012

L'analyse canonique entre le premier et le deuxième groupe donne une première corrélation canonique égal à 0.636, celle des groupes 1 et 3 donne une première corrélation canonique égale à 0.592 et enfin, celle des groupes 2 et 3 donne une premire corrélation canonique égale à 0.552. Dans les tableaux 8, 9 et 10, nous avons donné les résultats obtenus avec ces 3 méthodes d'ACG (méthodes ECART, MINVAR et MAXVAR). Le tableau 8 donne les vecteurs canoniques d'ordre un obtenus avec les 3 méthodes; le tableau 9 donne les valeurs propres ainsi que la différence entre la plus grande et la plus petite valeur propre de Φ et enfin, le tableau 10 donne les corrélations entre les variables canoniques d'ordre 1.

TABLEAU 8
Vecteurs canoniques d'ordre un

Méthodes	Groupe 1			Groupe 2			Groupe 3		
	X_1^1	X_2^1	X_3^1	X_1^2	X_2^2	X_3^2	X_1^3	X_3^1	X_3^3
MAXVAR	0.611	0.098	0.785	0.975	-0.222	-0.018	0.746	-0.514	0.421
MINVAR	0.902	-0.429	0.017	-0.012	0.472	-0.881	-0.909	-0.416	-0.012
ECART	0.064	-0.246	0.966	0.947	-0.315	-0.044	0.171	-0.413	0.894

TABLEAU 9
Les valeurs propres de Φ

Méthodes	$\lambda_1(\Phi)$	$\lambda_2(\Phi)$	$\lambda_3(\Phi)$	$\lambda_1(\Phi) - \lambda_3(\Phi)$
MAXVAR	1.8166	0.7274	0.4558	1.3607
MINVAR	1.5685	1.4301	0.0012	1.5672
ECART	1.7132	1.2706	0.0160	1.6972

TABLEAU 10
Corrélations entre les variables canoniques d'ordre 1

Méthodes	MAXVAR			MINVAR			ECART		
	Gr[1]	Gr[2]	Gr[3]	Gr[1]	Gr[2]	Gr[3]	Gr[1]	Gr[2]	Gr[3]
Gr[1]	1.000	0.528	0.292	1.000	-0.442	-0.562	1.000	0.631	-0.273
Gr[2]	0.528	1.000	0.395	-0.442	1.000	-0.490	0.631	1.000	0.550
Gr[3]	0.292	0.395	1.000	-0.562	-0.490	1.000	-0.273	0.550	1.000

Pour la table 8, nous constatons que :

(i) pour le premier groupe de variables, le poids de la deuxième variable obtenu avec la méthode MAXVAR (*cf* ligne 1 du tableau 8), de la première variable obtenu avec la méthode ECART (*cf.* ligne 3 du tableau 8) ainsi que, le poids de la troisième variable obtenu avec la méthode MINVAR (*cf.* ligne 2 du tableau 8) sont proches de 0. Ceci conduit à la conclusion suivante : la contribution de ces 3 variables à la variable canonique associée est faible.

(ii) pour le deuxième groupe, le poids de la troisième variable (*cf* MAXVAR et ECART) et celui de la première variable (*cf.* MINVAR) sont proches de 0.

(iii) pour le troisième groupe de variables, les vecteurs canoniques sont tous différents quel que soit la méthode utilisée. Toutes les variables contribuent à la formation de la variable canonique associée exceptée la troisième variable qui à un poids proche de 0 (*cf.* la méthode MINVAR).

L'analyse des résultats de la table 9 montre que, les matrices Φ de ces méthodes d'ACG n'ont pas les mêmes valeurs propres.

La valeur optimale de la méthode MAXVAR est égale à 1.817 (la plus grande valeur propre de Σ), celle de la méthode MINVAR est égale à 0.0012 (la plus petite valeur propre de Σ) et enfin, celle de la méthode ECART est égale à 1.697.

En ce qui concerne les corrélations linéaires entre les variables canoniques d'ordre un obtenues avec les 3 méthodes (*cf* tableau 10), on a :

- pour la méthode MAXVAR, les variables canoniques des groupes 1 et 2 sont les seuls qui sont corrélés entre eux. Leur corrélation est égale à 0.528.

- Pour la méthode MINVAR, les variables canoniques des 3 groupes de variables ont une corrélation deux à deux élevée en valeur absolue (corrélation proche de 0.5). Et enfin,
- pour la méthode ECART, les variables canoniques des groupes 1 et 2 (respectivement des groupes 2 et 3) sont aussi corrélées.

Conclusion : Pour ce deuxième exemple, les méthodes donnent des éclairages différents en ce qui concerne les liaisons linéaires entre les groupes de variables. La méthode qui donnent des résultats attendus (les trois groupes de variables sont corrélées) est la méthode MINVAR suivies, de la méthode ECART (les groupes 1 et 2 (respectivement 2 et 3) sont corrélées). La méthode MAXVAR quant à elle, montre que les seuls groupes de variables liés entre eux sont les groupes 1 et 2. Ceci contredit les résultats obtenus en faisant l'analyse canonique des groupes pris 2 à 2.

3.2.3. Troisième exemple

Les données utilisées dans cet exemple ont été publiées et analysées dans [Chessel 98] et dans [Escofier et Pagès 88]. Il s'agit d'un tableau de données formé de 21 individus (vins) et de 27 variables (variables sensorielles). Les variables forment 4 groupes associées aux 4 périodes de la dégustation du vin.

- Le premier groupe décrit 5 variables relatives à l'olfaction (vin au repos) : RInten (Intensité globale de l'arôme du vin au repos dans le verre), RQual (Qualité globale de l'arôme), RFruit (Note fruitée), RFleur (Note fleurie), REpice (Note épicée).
- Le deuxième groupe utilise trois variables associées à la vision : VInten (Intensité colorée), VNuance (Nuance violacée (Orangés= 0, à violet= 5)), VSurf (Impression de surface (traces laissées sur le verre)).
- Le troisième groupe décrit à partir de 10 variables, l'arÙme du vin après agitation et dégustation : AInten (Intensité de l'arÙme du vin dans le verre), AQuali (Qualité de l'arÙme du vin dans le verre), AFruit (Note fruitée), AFleur (Note fleurie), AEpice (Note épicée), AVeg (Note végétale), Aphenol (Note phénolique), ANoteb (Intensité globale de l'arÙme du vin dans la bouche), APersi (Persistance aromatique), AQual (Qualité des l'arÙme du vin dans la bouche).
- Le quatrième groupe contient 9 variables et décrit les qualités gustatives du vin : GInten (Intensité d'attaque), GAcid (Acidité en bouche), GAstr (Astringence en bouche), Galcool (Chaleur en bouche (alcool)), GEqui (Equilibre entre dominantes), GVelou (Velouté), GAmer (Amertume), GIfin (Intensité de fin en bouche), GHarmo (Structure-harmonique).

Les tableaux 11, 12, 13 et 14 donnent respectivement les valeurs propres de Σ, les corrélations linéaires entre les groupes de variables, les valeurs propres ainsi que la différence entre la plus grande et la plus petite valeur propre de Φ et enfin, les corrélations linéaires entre les variables initiales et les variables canoniques (ces derniers résultats correspondent aux figures 1 et 2).

TABLEAU 11
Les valeurs propres de la matrice des corrélations des 27 variables (matrice Σ)

$\lambda_1(\Sigma)$	$\lambda_2(\Sigma)$	$\lambda_3(\Sigma)$	$\lambda_4(\Sigma)$	$\lambda_5(\Sigma)$	$\lambda_6(\Sigma)$	$\lambda_7(\Sigma)$	$\lambda_8(\Sigma)$	$\lambda_9(\Sigma)$	$\lambda_{10}(\Sigma)$
3.69	3.24	2.49	2.27	2.17	1.87	1.85	1.61	1.52	1.28
$\lambda_{11}(\Sigma)$	$\lambda_{12}(\Sigma)$	$\lambda_{13}(\Sigma)$	$\lambda_{14}(\Sigma)$	$\lambda_{15}(\Sigma)$	$\lambda_{16}(\Sigma)$	$\lambda_{17}(\Sigma)$	$\lambda_{18}(\Sigma)$	$\lambda_{19}(\Sigma)$	$\lambda_{20}(\Sigma)$
1.09	0.99	0.85	0.69	0.48	0.44	0.21	0.09	0.07	0.03
$\lambda_{21}(\Sigma)$	$\lambda_{22}(\Sigma)$	$\lambda_{23}(\Sigma)$	$\lambda_{24}(\Sigma)$	$\lambda_{25}(\Sigma)$	$\lambda_{26}(\Sigma)$	$\lambda_{27}(\Sigma)$			
0.00	0.00	0.00	0.00	0.00	0.00	0.00			

TABLEAU 12
Corrélations entre les variables canoniques

Méthodes		Solution d'ordre 1 (Φ)				Solution d'ordre 2 (Φ)			
		Gr[1]	Gr[2]	Gr[3]	Gr[4]	Gr[1]	Gr[2]	Gr[3]	Gr[4]
ECART	Gr[1]	1.000	0.806	0.874	0.824	1.000	0.494	0.912	0.949
	Gr[2]	0.806	1.000	0.940	0.942	0.494	1.000	0.496	0.630
	Gr[3]	0.874	0.940	1.000	0.984	0.912	0.496	1.000	0.902
	Gr[4]	0.824	0.942	0.984	1.000	0.949	0.630	0.902	1.000
MAXVAR	Gr[1]	1.000	0.807	0.883	0.826	1.000	0.498	0.917	0.933
	Gr[2]	0.807	1.000	0.942	0.944	0.498	1.000	0.496	0.644
	Gr[3]	0.883	0.942	1.000	0.972	0.917	0.496	1.000	0.906
	Gr[4]	0.826	0.944	0.972	1.000	0.933	0.644	0.906	1.000
MINVAR	Gr[1]	1.000	0.488	-0.727	-0.325	1.000	-0.065	-0.664	0.202
	Gr[2]	0.488	1.000	-0.446	-0.609	-0.065	1.000	-0.171	-0.274
	Gr[3]	-0.727	-0.446	1.000	-0.198	-0.664	-0.171	1.000	-0.697
	Gr[4]	-0.325	-0.609	-0.198	1.000	0.202	-0.274	-0.697	1.000

TABLEAU 13
Les valeurs propres de la matrice Φ. $\lambda(\Phi) = \lambda_{max}(\Phi) - \lambda_{min}(\Phi)$

	Solution d'ordre 1			Solution d'ordre 2		
Méthode	$\lambda_{max}(\Phi)$	$\lambda_{min}(\Phi)$	$\lambda(\Phi)$	$\lambda_{max}(\Phi)$	$\lambda_{min}(\Phi)$	$\lambda(\Phi)$
MAXVAR	3.692	0.020	3.671	3.239	0.048	3.191
ECART	3.688	0.011	3.677	3.234	0.036	3.197
MINVAR	2.251	0.030	2.221	1.714	0.001	1.713

TABLEAU 14
Corrélations entre les variables initiales et les variables canoniques des 4 groupes

Groupe		Méthode ECART		Méthode MAXVAR		Méthode MINVAR	
	Variables	Ordre 1	Ordre 2	Ordre 1	Ordre 2	Ordre 1	Ordre2
Groupe 1	Rinten	0.629	-0.640	0.636	-0.637	0.110	-0.585
	Rqual	0.937	0.109	0.939	0.123	0.430	-0.156
	Rfruit	0.789	0.303	0.781	0.336	0.347	-0.030
	Rfleur	0.481	0.162	0.475	0.139	0.911	-0.313
	Repice	0.002	-0.910	0.011	-0.885	-0.152	0.136
Groupe 2	Vinten	0.883	-0.469	0.884	-0.466	0.759	0.457
	Vnuance	0.866	-0.482	0.866	-0.477	0.738	0.579
	Vsurf	0.998	-0.006	0.997	-0.002	0.972	0.215
Groupe3	Ainten	0.594	-0.577	0.617	-0.573	-0.362	0.738
	Aquali	0.814	0.432	0.807	0.433	-0.357	-0.198
	Afruit	0.725	0.480	0.740	0.487	-0.379	-0.124
	Afleur	0.191	0.307	0.176	0.300	-0.667	0.031
	Aepice	0.286	-0.665	0.272	-0.653	-0.021	-0.118
	Aveg	-0.531	-0.476	-0.512	-0.480	0.528	0.430
	Aphenol	0.401	-0.193	0.393	-0.190	0.054	0.355
	Anoteb	0.929	0.062	0.941	0.064	-0.376	0.274
	Apersi	0.937	-0.172	0.935	-0.171	-0.675	0.178
	Aqual	0.755	0.457	0.748	0.458	-0.632	-0.366

TABLEAU 14 (suite)
Corrélations entre les variables initiales et les variables canoniques des 4 groupes

Groupe		Méthode ECART		Méthode MAXVAR		Méthode MINVAR	
	Variables	Ordre 1	Ordre 2	Ordre 1	Ordre 2	Ordre 1	Ordre2
Groupe4	Ginten	0.821	-0.214	0.824	-0.205	-0.080	-0.355
	Gacid	-0.164	-0.187	-0.170	-0.189	0.371	0.182
	Gastr	0.793	-0.551	0.808	-0.530	-0.407	-0.636
	Galcool	0.816	-0.044	0.815	-0.027	-0.449	-0.305
	Gequi	0.787	0.388	0.782	0.407	-0.375	0.121
	Gvelou	0.827	0.265	0.829	0.272	-0.392	0.024
	Gamer	0.405	-0.565	0.410	-0.543	-0.098	-0.701
	Gifin	0.933	-0.148	0.930	-0.141	-0.291	0.271
	Gharmo	0.905	0.078	0.901	0.083	-0.419	0.133

En analysant les résultats du tableau 12, nous constatons que

- Les corrélations entre les variables canoniques sont presque identiques pour les méthodes MAXVAR et ECART.
- Pour les méthodes MAXVAR et ECART, les liaisons entre les 4 groupes de variables sont très élevées (car les inter-corrélations entre les variables canoniques d'ordre 1 (respectivement d'ordre 2) sont supérieures à 0.8 *cf* tableau 12, colonnes 1 à 4 et (respectivement sont supérieures à 0.49 *cf* tableau 12, colonnes 5 à 8)). C'est-à-dire qu'il existe bien une forte liaison linéaire entre les 4 groupes de variables.
- Pour la méthode MINVAR, les liaisons entre les 4 groupes de variables ne sont pas très élevées en général. Pour les variables canoniques d'ordre 1, les couples des variables canoniques qui ont une corrélation élevée sont : le couple des variables canoniques associées aux groupes 1 et 3 et, le couple des variables canoniques associées aux groupes 2 et 4.

En ce qui concerne les résultats du tableau 14, nous constatons que :

- Les variables initiales qui sont fortement corrélées avec la variable canonique d'ordre 1 obtenue avec les méthodes MAXVAR et ECART sont :

(1) RInten (Intensité globale de l'arôme du vin au repos dans le verre), RQual (Qualité globale de l'arôme), RFruit (Note fruitée) et RFleur (Note fleurie) pour **le groupe 1**(1)

(2) VInten (Intensité colorée), VNuance (Nuance violacée (Orangés = 0, à violet = 5)) et VSurf (Impression de surface (traces laissées sur le verre)) pour **le groupe 2**.

(3) AInten (Intensité de l'arôme du vin dans le verre), AQuali (Qualité de l'arôme du vin dans le verre), AFruit (Note fruitée) AVeg (Note végétale), Aphenol (Note phénolique), ANoteb (Intensité globale de l'arôme du vin dans la bouche), APersi (Persistance aromatique) et AQual (Qualité de l'arôme du vin dans la bouche) pour **le groupe 3**.

(4) et enfin, GInten (Intensité d'attaque), Gastr (Astringence en bouche), Galcool (Chaleur en bouche (alcool)), GEqui (Equilibre entre dominantes), GVelou (Velouté), GAmer (Amertume), GIfin (Intensité de fin en bouche) et GHarmo (Structure-harmonique) pour **le groupe 4**.

- Pour la méthode MINVAR, les variables initiales qui sont fortement corrélées avec sa première variable canonique sont :

  (1) RFleur (Note fleurie) pour **le groupe 1**.

  (2) VInten (Intensité colorée), VNuance (Nuance violacée (orangés = 0, à violet = 5)) et VSurf (Impression de surface (traces laissées sur le verre)) pour **le groupe 2**.

  (3) AFleur (Note fleurie), AVeg (Note végétale), APersi (Persistance aromatique) et AQual (Qualité de l'arÙme du vin dans la bouche) pour **le groupe 3** et enfin,

  (4) aucune variable du **groupe 4** n'est bien liée avec sa variable canonique.

En ce qui concerne les corrélations entre les variables initiales et les variables canoniques d'ordre deux, nous constatons que, pour les trois méthodes, la plupart des variables initiales ne sont pas fortement corrélées avec la variable canonique d'ordre deux.

Pour les méthodes MAXVAR et ECART, les variables qui ont une forte corrélation, en valeur absolue, avec cet axe sont : les variables Rinten, Repice, Ainten, Afruit, Aepice, Gastr et Gamer.

Pour les méthodes MINVAR, les variables qui ont une forte corrélation, en valeur absolue, avec la deuxième variable canonique sont : les variables Rinten, Vnuance, Ainten, Gastr et Gamer.

Conclusion : Pour cet exemple, les méthodes donnent aussi des résultats différents. Les méthodes ECART et MAXVAR montrent qu'il existe une forte liaison linéaire entre les trois groupes de variables tandis que, pour la méthode MINVAR, les groupes de variables ne sont pas globalement liés entre eux, bien que la plus grande valeur propre de Σ est proche de 4 ($Im(X^1) \cap Im(X^2) \cap Im(X^3) \cap Im(X^4) \neq \{vide\}$ pour tout i et j).

Les figures 1 et 2 (respectivement 3 et 4) donnent, pour chaque groupe de variables et pour chacune des 3 méthodes, les représentations graphiques des variables dans le plan factoriel engendré par les deux premières variables canoniques (respectivement les représentations graphiques des 21 vins dans le plan factoriel engendré par les variables canoniques d'ordre 1 et 2).

Les figures 1 et 2 (respectivement 3 et 4) montrent que les 3 méthodes d'ACG présentées ici ne donnent pas en général les mêmes résultats en ce qui concerne l'étude des liaisons linéaires entre les variables canoniques et les variables initiales

(respectivement l'étude du comportement des 21 vins). D'un groupe de variables à l'autre, la dispersion des 21 vins n'est pas similaire.

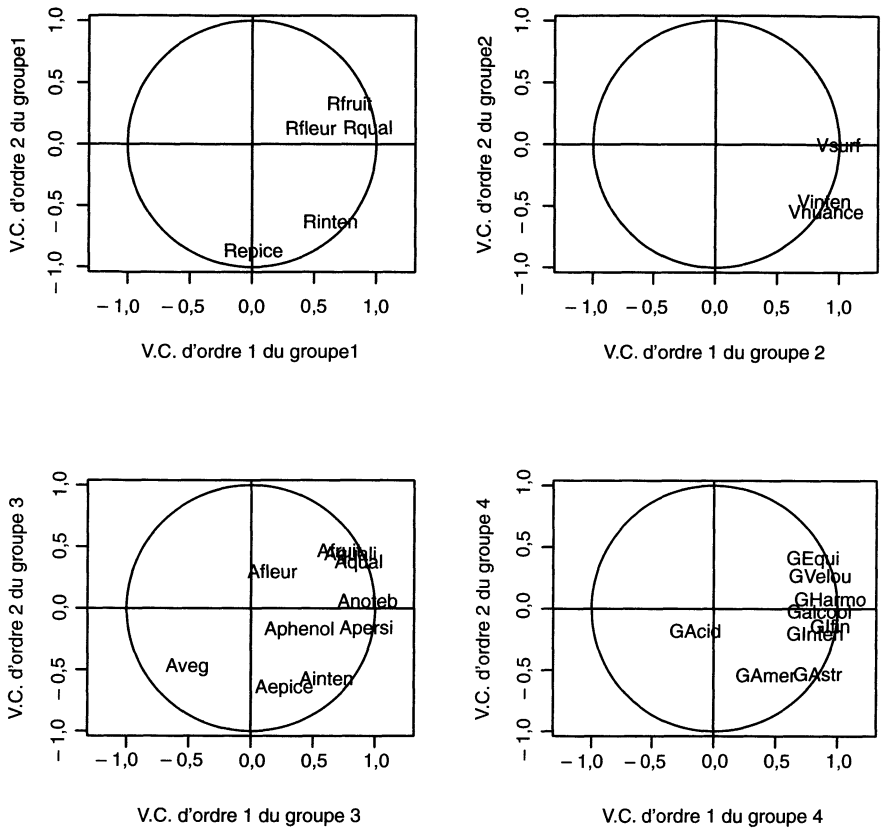

FIGURE 1

Méthodes MAXVART et ECART. Représentation graphique, pour chaque groupe, des variables dans le plan factoriel engendré par les deux premières variables canoniques (ici, nous avons donné une seule figure pour les deux méthodes, car elles donnent des variables canoniques pratiquement identiques)

Ceci met en évidence l'influence du critère sur les solutions des méthodes d'analyse canonique généralisée et, de l'importance du choix de la méthode d'ACG pour étudier les liaisons linéaires entre plusieurs groupes de variables.

On remarque, une très bonne représentation de l'information disponible sur ces 4 cercles de corrélations ci-dessous. Pour les groupes 1, 2 et 4, toutes les variables sont bien représentées excepté la variable Rfleur (pour le groupe 1) et les variables Gacid et Gamer (pour le groupe 4). Pour le groupe 3, nous remarquons que les variables Afleur et Aphenol sont mal représentées sur le cercle de corrélation.

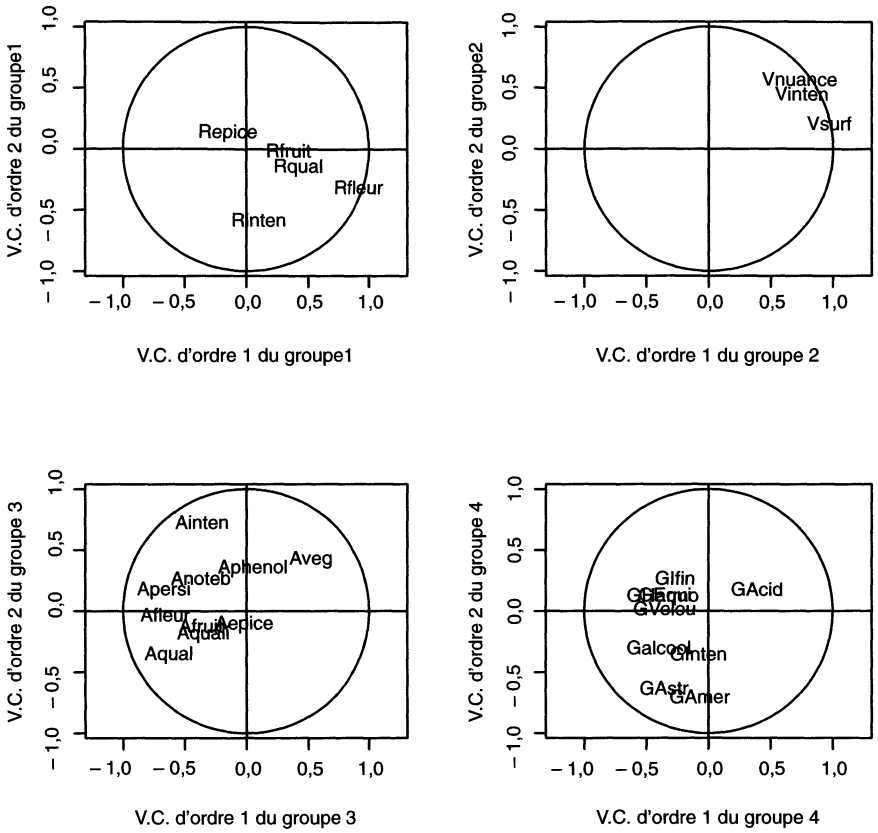

FIGURE 2
Méthode MINVAR. Représentation graphique, pour chaque groupe, des variables dans le plan factoriel engendré par les deux premières variables canoniques

Pour la méthode MINVAR, on remarque une mauvaise représentation de l'information disponible sur 3 des 4 cercles de corrélations ci-dessous. En effet, pour les groupes 1, 3 et 4 la plupart des variables sont mal représentées sur ces cercles de corrélations excepté la variable Rfleur (pour le groupe 1) et les variables Ainten, Apersi, Afleur, Aqual et Aveg (pour le groupe 3). Pour le groupe 2, toutes les variables sont bien représentées.

4. Conclusion

Dans cet article, nous avons proposé une méthode d'ACG appelée méthode ECART. La méthode permet d'étudier les liaisons linéaires, existante, entre plusieurs groupes de variables. Elle se réduit à l'analyse canonique linéaire proposée par Hotelling dans le cas de deux groupes de variables. Elle maximise une mesure

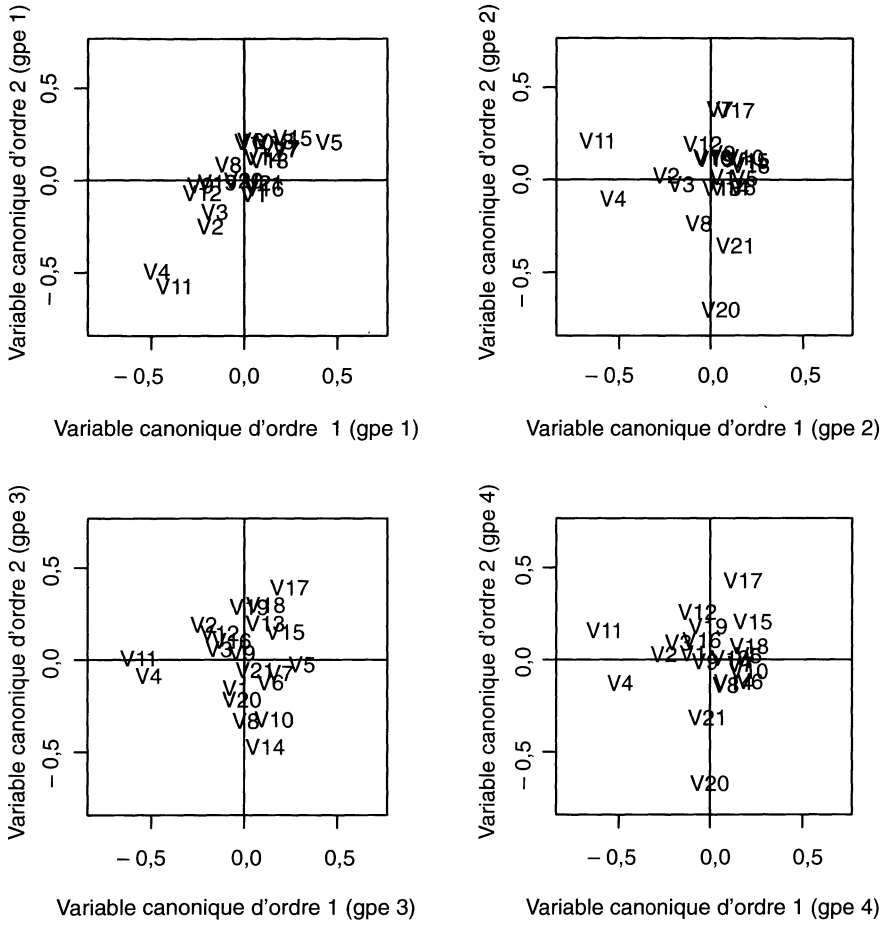

FIGURE 3

Méthodes MAXVAR et ECART : Représentation graphique du nuage des points des 21 vins, numérotés de V1 à V21, dans le plan factoriel Ici, nous avons donné une seule figure pour les 2 méthodes, car elles donnent des variables canoniques pratiquement identiques

d'association généralisée qui est nulle si et seulement si toutes les valeurs propres de Φ sont égales (c'est-à-dire quand les variables canoniques sont deux à deux orthogonales) et elle est égale à 1, quand toutes les valeurs propres de Φ sont égales à 0 excepté la première qui est égale à p (le nombre de groupes). Les figures 1, 2, 3 et 4 (la représentation graphique des variables et la représentation graphique des individus dans les différents plans engendrés par les variables canoniques) associées s'interprètent comme ceux des autres méthodes d'ACG.

Après avoir décrit cette méthode, nous l'avons comparée à deux autres méthodes d'ACG : la méthode MAXVAR et la méthode MINVAR. La comparaison de ces trois

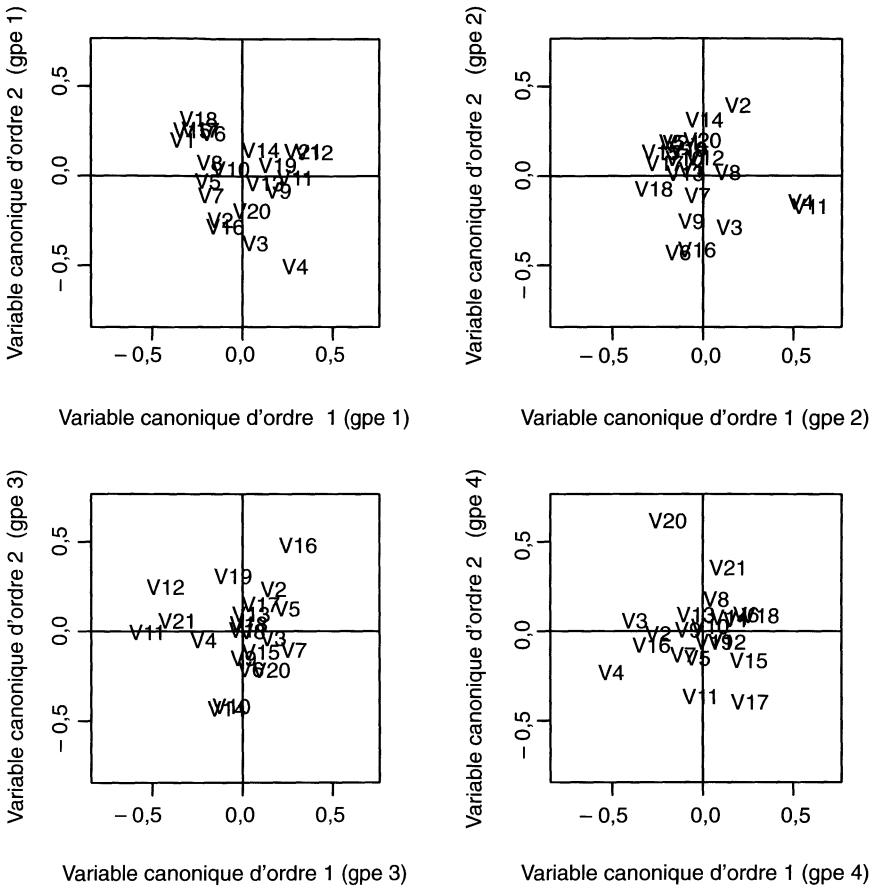

FIGURE 4

Méthode MINVAR : Représentation graphique du nuage des points des 21 vins, numérotés de V1 à V21, dans le plan factoriel engendré par les deux premières variables canoniques

méthodes d'ACG a été faite de deux façons. La première était théorique et la seconde numérique. Trois exemples ont été traités pour comparer les trois méthodes d'ACG.

Pour le premier exemple, les 3 méthodes conduisent à la même conclusion, il y'a aucune liaison linéaire entre les groupes de variables. Ces résultats étaient attendus dans la mesure où il n'y avait pas de liaisons linéaires entre les groupes de variables pris 2 à 2. Pour le deuxième exemple, les 3 méthodes donnent des éclairages différents (les corrélations linéaires entre les groupes de variables sont différentes d'une méthode d'ACG à l'autre) et enfin, pour le troisième exemple, les méthodes ECART et MAXVAR conduisent aux mêmes conclusions : Il y'a une forte liaison linéaire entre les groupes de variables. Ces résultats étaient aussi attendus puisque la plus grande valeur propre de Σ était proche de 4 (le nombre total de groupe de

variables) tandis que la dernière valeur propre était égale à 0. Pour cet exemple, la méthode MINVAR donne des résultats différents de ceux obtenus avec les méthodes ECART et MAXVAR. En ce qui concerne l'algorithme S que nous avons proposé, ce dernier peut être aussi utilisé pour calculer les solutions de MAXVAR et de MINVAR.

La source SPLUS du programme de la méthode ECART peut être obtenue sur simple demande chez les auteurs.

Remerciements

Nous remercions vivement le professeur P. Cazes pour sa lecture attentive et pour les remarques qui ont contribué à l'amélioration de cet article.

Bibliographie

[Anderson 84] ANDERSON T.W. (1984), An introduction to multivariate analysis, 2nd edition, J.Wiley.

[Chessel 98] CHESSEL D. (1998), L'Analyse factorielle multiple. ADE-4/Fiche thématique 5.3/97-07.

[Escofier et Pagès 88] ESCOFIER B. et PAGÈS J. (1988), Analyses Factorielles Simples et Multiples : objectifs, méthodes et interprétations. Dunod.

[Foucart 96] FOUCART T. (1996), Analyse de la colinéarité classification des variables. Rev. Statistique Appliquée, XLIV, pp. 41-57.

[Hotelling 36] HOTELLING H. (1936), Relations between two sets variables. Biometrika vol 28, pp. 321-377.

[Horst 61a] HORST P. (1961), Relations among m sets of measures. Psychometrika, vol.26, no.2, pp. 129-149.

[Horst 61b] HORST P. (1961), Generalized canonical correlations and their applications to experimental data. J. Clinical psychol. vol.14, pp. 331-347.

[Kettenring 71] KETTENRING J.R. (1971), Canonical analysis of several sets of variables. Biometrika, 58, 3, pp. 433-451.

[Nzobounsana 01] NZOBOUNSANA V. (2001), L'analyse canonique généralisée : méthodes, applications et sélection des variables dans les groupes. Thèse soutenue à l'université de Rennes 2.

[Nzobounsana et Dhorne 02] NZOBOUNSANA V. et DHORNE T. (2002), Influence du critère sur les solutions des méthodes d'analyse canonique généralisée. Communication au xxxiv Journées de Statistique. Bruxelles et Louvain-la-Neuve 2002.