Rev. Statistique Appliquée, 2006, **LIV** (1), 27-45

CLASSIFICATION DE VARIABLES AUTOUR DE COMPOSANTES LATENTES

E. VIGNEAU[(1)], E.M. QANNARI[(1)], K. SAHMER[(1)], D. LADIRAY[(2)]

[(1)] *ENITIAA / INRA, Unité de Sensométrie et de Chimiométrie, Nantes, France*
[(2)] *INSEE, Département des Comptes Nationaux, Malakoff, France*

RÉSUMÉ

Une approche de classification de variables autour de composantes latentes, nommée CLV, est proposée. Cette approche englobe différents cas de figures : le cas où l'utilisateur souhaite grouper des variables corrélées entre elles sans tenir compte du signe de la corrélation et le cas où une corrélation négative traduit une opposition entre variables. L'approche offre également la possibilité de définir les variables latentes associées aux groupes comme étant des combinaisons linéaires de variables externes. La classification de variables s'intègre ainsi dans des contextes variés : analyse en composantes principales, régression linéaire multiple, régression PLS ou analyse en composantes principales sur variables instrumentales. Les algorithmes d'optimisation des critères qui sous-tendent la méthode de classification de variables sont décrits et une comparaison avec la procédure de classification de variables, Varclus, implémentée dans le logiciel SAS est proposée.

Mots-clés : *Classification de variables, analyse en composantes principales, régression PLS*

ABSTRACT

An approach for clustering variables around latent components, named CLV, is presented. This method aims at clustering variables in two situations : the case where variables are to be grouped according to the magnitude of their correlation without taking account of the sign of the correlation and the case where negatively correlated variables are considered as dissimilar. The possibility of taking account of external variables by expressing the latent variables associated with the various clusters as linear combinations of these external variables is also included. Thus CLV may be involved in several contexts such as principal components analysis, multiple linear regression, PLS regression or redundancy analysis. Algorithms for the optimisation of the criteria that underly CLV analysis as well as a comparison with the procedure Varclus implemented in SAS software are discussed.

Keywords : *Clustering of variables, Principal Components Analysis, PLS Regression*

1. Introduction

La plupart des méthodes de classification sont conçues pour classer des individus. Cependant, dans d'autres types d'application, l'intérêt porte sur la classification des variables utilisées pour décrire les objets. La segmentation d'un panel de consommateurs, lorsque les consommateurs ont fourni des notes de préférence pour un ensemble de produits, est une application typique dans laquelle la question se pose en terme de classification de variables (ici les consommateurs). La classification de variables économiques peut être mise en œuvre afin d'identifier certains types de stratégies financières, l'accent étant mis sur la typologie des stratégies plutôt que sur la typologie des unités adoptant la même stratégie.

En tant qu'outil d'investigation de la structure d'un tableau de données, la classification de variables peut être vue comme une approche complémentaire de l'analyse factorielle. De plus, en permettant l'identification de groupes de variables homogènes, la classification de variables permet d'exhiber de nouvelles directions faciles à interpréter, à l'instar des techniques de rotation orthogonale ou oblique des facteurs, telles que Varimax, Promax, ... (Harman, 1976). Cependant contrairement aux techniques de rotation, la classification directe des variables aboutit à une classification en sous-ensembles disjoints, ce qui rend l'interprétation des structures sous-jacentes plus facile. La classification de variables, enfin, est une étape intermédiaire possible dans une stratégie de sélection de variables. Cette approche est une alternative aux techniques de sélection de variables développées par exemple par Jolliffe (1972), McCabe (1984), Krzanowski (1987), Al-Kandari et Jolliffe (2001) ou Guo *et al*. (2002).

Dans la plupart des références abordant le problème de la classification de variables, l'approche repose sur la définition d'un indice de similarité ou de dissimilarité entre variables (Derquenne, 1997; Qannari *et al.*, 1998; Abdallah et Saporta, 1998; Soffritti, 1999). Le travail présenté ici repose sur un principe différent qui consiste à définir des groupes de variables autour de composantes latentes de sorte que chaque variable soit liée de manière optimale à la composante du groupe auquel elle est affectée. La recherche d'une variable représentative (ou composante latente) dans chaque groupe fait partie intégrante de la démarche. Un lien peut être établi entre ces composantes latentes et les composantes factorielles en Analyse en Composantes Principales (ACP) ou en régression PLS (PLSR). Vigneau et Qannari (2003b) ont montré la complémentarité de ces approches dans le cadre de l'analyse sensorielle.

La méthode de classification de variables (CLV) que nous proposons peut être mise en œuvre dans différents contextes et s'adapte en fonction de l'objectif visé. Pour cela, deux paramètres ont été introduits. Le premier paramètre est relatif à la nature des groupes de variables recherchés. L'objectif peut être :

– de regrouper des variables ayant des coefficients de corrélation ou de covariance importants, sans tenir compte du signe de ces coefficients. Dans cette situation, chaque groupe de variables est structuré autour d'un axe représenté par la composante latente du groupe;

– de considérer que deux variables corrélées négativement sont « éloignées » et donc de tenir compte du signe de la corrélation entre variables. Dans ce cas, un groupe de variables est défini localement.

FIGURE 1
Deux cas de figure pris en compte par la méthode CLV. Le premier cas correspond au cas où l'importance de la relation linéaire entre variables, mais non son signe, est prise en considération. Le deuxième cas correspond au cas où une corrélation négative entre variables traduit un désaccord entre ces variables

Ces deux optiques sont schématisées dans la figure 1. Il est à noter que la procédure Varclus implémentée dans le logiciel SAS/STAT, dédiée à la classification de variables intègre également ces deux options. Une comparaison plus précise de la procédure Varclus et de CLV sera abordée par la suite.

Le second paramètre intégré à la méthode CLV est la prise en compte, ou non, de variables externes. Lorsque des variables externes sont disponibles, en plus des variables à grouper, il peut être souhaitable d'imposer aux composantes latentes des groupes d'être des combinaisons linéaires de ces variables externes. L'intérêt est alors de pouvoir interpréter les structures exhibées dans les groupes à partir de ces informations complémentaires. Le contexte de la cartographie externe des préférences d'un panel de consommateurs (Greenhoff et MacFie, 1994) illustre très bien cette situation. Dans ce type d'étude, l'objectif est d'identifier des groupes de consommateurs ayant exprimé de manière similaire leur préférence pour un ensemble de produits et d'expliquer ces préférences à l'aide des caractéristiques sensorielles ou physico-chimiques des produits, ces dernières variables étant les variables externes.

La méthode CLV que nous proposons apparaît, de fait, duale des méthodes d'analyse factorielle typologique ou de régressions locales présentées par Diday (1974, 1976). En fonction de l'objectif visé, la démarche proposée par cet auteur consiste, en effet, à déterminer des groupes d'individus de sorte à optimiser dans chaque classe la représentation des individus selon des axes factoriels, des axes discriminants ou encore à optimiser l'ajustement à des modèles de régression linéaire intra-classes.

La partie suivante est consacrée à la présentation de la méthode CLV. Les critères et algorithmes sur lesquels reposent la méthode seront explicités en fonction

des différents cas de figure évoqués ci-dessus. Une dernière partie sera consacrée à deux études de cas.

2. Méthode

2.1. Classification de variables lorsque le signe des corrélations n'est pas pertinent

Considérons tout d'abord le cas où seule la force de la liaison linéaire entre deux variables importe, sans prendre en considération le sens de cette liaison. L'objectif est donc de définir K groupes de variables (K étant fixé) et K composantes latentes de sorte à optimiser un critère qui traduit le degré de liaison linéaire entre les variables d'un groupe et la composante associée à ce groupe.

Soient $\mathbf{x_1}$, $\mathbf{x_2}$, ..., $\mathbf{x_p}$ les p variables à classer. Ces variables sont mesurées sur n individus et sont supposées être centrées. Nous cherchons K groupes de variables, G_1, G_2, ... ,G_K et K composantes latentes, $\mathbf{c_1}$, $\mathbf{c_2}$, ..., $\mathbf{c_K}$, associées respectivement à ces groupes de manière à minimiser le critère :

$$E = \frac{1}{n} \sum_{k=1}^{K} \sum_{j=1}^{p} \delta_{kj} \, \|\mathbf{x_j} - \alpha_{\mathrm{kj}} \mathbf{c_k}\|^{\mathbf{2}} \tag{1}$$

où $\delta_{kj} = 1$ si la variable $\mathbf{x_j}$ appartient au groupe G_k et $\delta_{kj} = 0$ sinon, et où α_{kj} est un réel à déterminer.

Pour une partition fixée, en développant le critère E et en considérant sa dérivée partielle par rapport à α_{kj}, il est facile de vérifier que les valeurs optimales de α_{kj} sont données par :

$$\alpha_{kj} = \frac{\mathbf{c_k^t x_j}}{\mathbf{c_k^t c_k}}$$

En substituant cette valeur dans l'expression de E, il vient que minimiser E équivaut à maximiser :

$$T = n \sum_{k=1}^{K} \sum_{j=1}^{p} \delta_{kj} \, Cov^2(\mathbf{x_j}, \mathbf{c_k^*}) = \frac{1}{n} \sum_{k=1}^{K} \mathbf{c_k^{*t} X_k X_k^t c_k^*} \text{ avec } \mathbf{c_k^*} = \frac{\mathbf{c_k}}{\sqrt{\mathbf{c_k^t c_k}}} \tag{2}$$

où $\mathbf{X_k}$ représente la matrice de données formée uniquement par les variables appartenant au groupe G_k.

Une solution au problème de classification considéré peut être obtenue en mettant en œuvre un algorithme de type nuées dynamiques (Diday, 1971). À partir d'une partition initiale en K groupes, on définit tout d'abord les composantes latentes des groupes. La composante latente normée $\mathbf{c_k^*}$ dans le groupe G_k est donnée par la

première composante principale normée de $\mathbf{X_k}$. Par la suite, les variables sont réaffectées dans les groupes en fonction du carré de leurs covariances avec chacune des composantes des groupes. Ainsi une variable $\mathbf{x_j}$ est affectée dans le groupe G_k si :

$$Cov^2(\mathbf{x_j}, \mathbf{c_k^*}) = \max_{g=1,\ldots,K} Cov^2(\mathbf{x_j}, \mathbf{c_g^*})$$

La procédure en deux étapes décrite ci-dessus est réitérée jusqu'à stabilisation de la partition des variables. On aboutit à la définition de composantes latentes dans les groupes comparables aux composantes principales de $\mathbf{X}$, la matrice complète de données, à la différence près qu'elles ne sont pas nécessairement deux à deux orthogonales. En contrepartie, elles sont plus faciles à interpréter car chaque composante de groupe est une combinaison linéaire d'un sous-ensemble de variables qui reflètent une même tendance.

Le problème de l'initialisation de l'algorithme de partitionnement est résolu dans la méthode CLV en mettant en œuvre, au préalable, un algorithme de classification ascendante hiérarchique fondé sur la maximisation du critère T (eq.2). La coupure du dendrogramme en K classes fournit une partition initiale quasi-optimale que l'algorithme de partitionnement va tenter de consolider.

À la première étape de l'algorithme hiérarchique, chaque variable forme un groupe à elle seule. Le critère T vaut alors :

$$T_1 = \sum_{j=1}^{p} var(\mathbf{x_j})$$

À une étape i donnée, $K_i = p - i + 1$ groupes de variables sont formés et la valeur optimale du critère T est :

$$T_i = \sum_{k=1}^{K_i} \lambda_1^{(k)} \tag{3}$$

où $\lambda_1^{(k)}$ est la première valeur propre de $\frac{1}{n}\mathbf{X_k}\mathbf{X_k^t}$.

Le passage de l'étape i à l'étape $(i+1)$ correspond à la fusion de deux groupes, disons A et B, ce qui résulte en une variation du critère T donnée par :

$$\Delta T = T_i - T_{i+1} = \lambda_1^{(A)} + \lambda_1^{(B)} - \lambda_1^{(A \cup B)}$$

On peut montrer que ΔT est positif ou nul, autrement dit que le critère T décroît d'une étape à une autre de la hiérarchie. La stratégie consiste par conséquent à agréger, à chaque étape, les deux groupes de variables conduisant à la plus petite diminution de T.

L'intérêt de l'algorithme hiérarchique est, non seulement d'obtenir une partition initiale pour l'algorithme de partitionnement, mais aussi de fournir une aide pour le choix du nombre K de groupes à retenir. En effet, l'évolution du critère d'agrégation

ΔT permet de détecter les niveaux d'agrégation auxquels la fusion de deux groupes se traduit par une forte diminution du critère T. Ceci indique des niveaux où le dendrogramme pourrait être coupé.

Comparativement à la méthode CLV, la procédure Varclus du logiciel SAS/-STAT présente de grandes similarités dans son principe : « la procédure Varclus a pour but de diviser un ensemble de variables en groupes distincts de sorte que chaque groupe puisse être interprété comme quasiment unidimensionnel » (SAS/STAT, 1990). À chaque groupe est associée une composante et la procédure Varclus vise à maximiser la somme des variances des composantes de groupe. Ce critère est identique au critère *T* donné par (2) ou (3). L'algorithme mis en œuvre par Varclus pour atteindre cet objectif est un algorithme hiérarchique descendant, alors que la construction d'une hiérarchie avec CLV est ascendante. À chaque étape, le groupe de variables le moins « unidimensionnel » est séparé en deux, l'unidimensionalité pouvant être évaluée soit par le pourcentage d'inertie expliqué par la première composante principale, soit par celui de la seconde composante principale. Les deux premières composantes principales du groupe à séparer, une fois transformées par une rotation de type quartimax, servent de support pour la division des variables en deux sous-ensembles : chaque variable est affectée à la composante avec laquelle sa corrélation au carré est la plus importante. Un second type de critère est utilisé et conduit à une ré-affectation éventuelle des variables : une variable peut changer de groupe si cela conduit à une augmentation de la variance expliquée par la composante du groupe. Cette seconde phase de consolidation est relativement complexe et demande de nombreux calculs. En comparaison, aucune ré-affectation n'est prévue au cours de la hiérarchie de la méthode CLV, ce qui se traduit par une programmation simplifiée. L'étape de ré-affectation est intégrée dans l'algorithme de partitionnement qui permet la consolidation des groupes et l'augmentation du critère T.

2.2. Extension de la classification de variables quand le signe des corrélations n'est pas pertinent, en présence de variables externes

L'utilisateur peut être intéressé par le regroupement de variables en tenant compte d'une information externe. Dans une perspective exploratoire, cette démarche permet d'expliquer une partition de variables à l'aide de variables connexes. Dans une perspective prédictive, cela permet d'élaborer des modèles de prédiction pour chacun des groupes de variables. La méthode CLV basée sur le critère E peut être aisément adaptée dans ce cas. Cette extension n'est pas envisageable avec la procédure Varclus de SAS.

Considérons un tableau $\mathbf{Z}$ constitué de q variables $\mathbf{z_1}$, $\mathbf{z_2}$, ..., $\mathbf{z_q}$ mesurées sur les mêmes individus, centrées et éventuellement réduites. Nous imposons maintenant, en plus du critère de cohérence interne des groupes, que la composante latente de chaque groupe soit combinaison linéaire des variables externes. Ainsi, on cherche à minimiser :

$$E^* = \frac{1}{n}\sum_{k=1}^{K}\sum_{j=1}^{p}\delta_{kj}\,\|\mathbf{x_j} - \alpha_{kj}\mathbf{c_k}\|^{\mathbf{2}} \quad \text{sous la contrainte} \quad \mathbf{c_k} = \mathbf{Z}\ \mathbf{b_k} \qquad (4)$$

Nous pouvons vérifier que ceci est équivalent au problème de maximisation de :

$$T^* = \frac{1}{n} \sum_{k=1}^{K} \frac{\mathbf{b_k^t}\, \mathbf{Z^t}\, \mathbf{X_k}\, \mathbf{X_k^t}\, \mathbf{Z}\, \mathbf{b_k}}{\mathbf{b_k^t}\, \mathbf{Z^t}\, \mathbf{Z}\, \mathbf{b_k}} \tag{5}$$

Pour une partition donnée en K_i groupes, l'optimum est donné par :

$$T_i^* = \sum_{k=1}^{K_i} \nu_1^{(k)}$$

où $\nu_1^{(k)}$ est la plus grande valeur propre de $\frac{1}{n}(\mathbf{Z^t}\, \mathbf{Z})^{-\mathbf{1}}\, \mathbf{Z^t}\, \mathbf{X_k}\, \mathbf{X_k^t}\, \mathbf{Z}$. Cet optimum est atteint pour le vecteur $\mathbf{b_k}$ qui est un vecteur propre associé à $\nu_1^{(k)}$. Il découle de ce résultat que la composante $\mathbf{c_k} = \mathbf{Z b_k}$, dans le groupe G_k, est la première composante de l'ACP sur Variables Instrumentales (ACPVI) du bloc des variables $\mathbf{X_k}$ sur le bloc des variables externes $\mathbf{Z}$ (Rao, 1964, Sabatier, 1987).

La résolution de ce problème se heurte cependant à une difficulté en présence de quasi-colinéarité entre les variables externes $\mathbf{Z}$. Du fait de l'inversion de la matrice $\mathbf{Z^t Z}$, l'incertitude associée à l'estimation des vecteurs de coefficients $\mathbf{b_k}$ peut être très grande. Il peut même s'avérer impossible d'estimer les vecteurs de coefficients $\mathbf{b_k}$ lorsque le nombre d'observations est inférieur au nombre de variables externes. Pour lever cette difficulté, une alternative est possible. Cette alternative, introduite antérieurement (Vigneau et Qannari, 2002, 2003a, 2003b), est une extension directe du critère T avec prise en compte des variables externes. Le nouveau critère considéré est :

$$\tilde{T} = n \sum_{k=1}^{K} \sum_{j=1}^{p} \delta_{kj}\, Cov^2(\mathbf{x_j}, \mathbf{c_k}) \text{ sous les contraintes } \mathbf{c_k} = \mathbf{Z}\ \mathbf{a_k} \text{ et } \mathbf{a_k}^{\mathrm{t}}\, \mathbf{a_k} = \mathbf{1} \tag{6}$$

Ce critère peut également s'écrire sous la forme :

$$\tilde{T} = \frac{1}{n} \sum_{k=1}^{K} \mathbf{a_k^t}\, \mathbf{Z^t}\, \mathbf{X_k}\, \mathbf{X_k^t}\, \mathbf{Z}\, \mathbf{a_k} \tag{7}$$

L'optimum pour $\tilde{T}$ lorsque la partition comporte K_i classes est :

$$\tilde{T_i} = \sum_{k=1}^{K_i} \mu_1^{(k)}$$

où $\mu_1^{(k)}$ est la plus grande valeur propre de $\frac{1}{n}\mathbf{Z^t}\, \mathbf{X_k}\, \mathbf{X_k^t}\, \mathbf{Z}$ et $\mathbf{a_k}$ est un vecteur propre associé à cette valeur propre. On vérifie qu'alors la composante $\mathbf{c_k} = \mathbf{Z a_k}$, dans le groupe G_k, est la première composante de la régression PLS2 du bloc des variables $\mathbf{X_k}$ sur le bloc des variables $\mathbf{Z}$.

Du point de vue de la mise en œuvre des calculs, les algorithmes hiérarchique et de partitionnement ont la même structure que dans le cas précédent (sans variable externe). La stratégie d'agrégation adoptée pour l'algorithme ascendant hiérarchique consiste à réunir, à chaque étape, les deux groupes de variables qui résultent en la plus petite diminution du critère considéré, T^* ou $\tilde{T}$. En effet, nous pouvons montrer que les critères T^* et $\tilde{T}$ décroissent entre deux étapes successives des procédures hiérarchiques.

2.3. Classification de variables lorsqu'une corrélation négative implique un désaccord

Dans certaines situations, il est souhaitable que chaque groupe de variables soit organisé localement, les variables d'un même groupe étant fortement et positivement corrélées.

La détermination de K groupes de variables, G_1, ..., G_K, repose cette fois-ci sur la minimisation du critère Q défini par :

$$Q = \frac{1}{n} \sum_{k=1}^{K} \sum_{j=1}^{p} \delta_{kj} \, \|\mathbf{x_j} - \mathbf{c_k}\|^{\mathbf{2}} \tag{8}$$

Le lien entre ce problème de classification et la classification d'individus dans un espace muni d'une distance euclidienne est aisé à faire.

Pour une partition donnée, le critère Q atteint son minimum pour :

$$\mathbf{c_k} = \bar{\mathbf{x}}_\mathbf{k}$$

où $\bar{\mathbf{x}}_\mathbf{k}$ représente la variable «centroïde» ou variable «moyenne» du groupe G_k.

À une étape i de l'algorithme ascendant hiérarchique, le critère Q vaut :

$$Q_i = \frac{1}{n} \sum_{k=1}^{K_i} \sum_{j=1}^{p} \delta_{kj} \, \|\mathbf{x_j}\|^2 - \frac{1}{n} \sum_{k=1}^{K_i} \sum_{j=1}^{p} \delta_{kj} \, \|\bar{\mathbf{x}}_\mathbf{k}\|^{\mathbf{2}} = \sum_{j=1}^{p} Var(\mathbf{x_j}) - \sum_{k=1}^{K_i} p_k \, Var(\bar{\mathbf{x}}_\mathbf{k}) \tag{9}$$

où p_k est le nombre de variables dans le groupe G_k.

En désignant par $W_i = \sum_{k=1}^{K_i} p_k \, Var(\bar{\mathbf{x}}_\mathbf{k})$, nous pouvons montrer que si de l'étape i à l'étape $(i+1)$ deux groupes de variables, disons *A* et *B*, sont fusionnés, le critère Q augmente, et par voie de conséquence, le critère W diminue d'une quantité :

$$\begin{aligned} \Delta W &= W_i - W_{i+1} \\ &= p_A Var(\bar{\mathbf{x}}_\mathbf{A}) + p_B Var(\bar{\mathbf{x}}_\mathbf{B}) - (p_A + p_B) \, Var(\bar{\mathbf{x}}_{\mathbf{A} \cup \mathbf{B}}) \\ &= \frac{p_A p_B}{p_A + p_B} Var(\bar{\mathbf{x}}_\mathbf{A} - \bar{\mathbf{x}}_\mathbf{B}) \end{aligned}$$

où $\bar{x}_A, \bar{x}_B, \bar{x}_{A\cup B}$ représentent les centroïdes respectifs des groupes A, B, A∪B et p_A (resp. p_B) le nombre de variables dans le groupe A (resp.B). Ainsi à chaque étape de la hiérarchie, les groupes agrégés sont ceux qui conduisent à la plus petite diminution de W (critère de Ward).

La consolidation de la partition obtenue par coupure de l'arbre hiérarchique à un niveau choisi par l'utilisateur est basée sur un algorithme de partitionnement autour des centres mobiles (Diday, 1971).

La procédure Varclus de SAS/STAT offre également la possibilité de construire une partition de variables autour de centroïdes définis comme la moyenne non pondérée des variables du groupe. Cependant, le critère d'unidimensionalité qui sert de base au choix du groupe à scinder dans Varclus n'est plus forcément pertinent lorsque l'on cherche des groupes «locaux» rassemblant des variables positivement corrélées. Un groupe peut être quasiment unidimensionel tout en rassemblant des variables négativement corrélées. Par ailleurs, la première phase de la procédure de ré-affectation des variables dans les groupes, utilisant le carré de la corrélation entre les variables et les «centroïdes» de groupe, est discutable. Pour cette raison, cette phase est limitée à une seule itération, au profit de la seconde phase de la procédure de ré-affectation (*search algorithm*) dont l'objectif est d'augmenter itérativement le niveau de la variance expliquée. En dépit de la différence entre les algorithmes, Varclus avec l'option CENTROID et CLV dans le cas présent sont basés sur le même critère d'évaluation de la partition, à savoir la somme pondérée des variances des variables «centroïdes» (eq.9).

L'avantage de la méthode CLV par rapport à Varclus réside dans la simplicité de mise en œuvre mais aussi dans la possibilité de tenir compte de variables externes comme cela est exposé dans le paragraphe suivant.

2.4. Extension de la classification de variables lorsqu'une corrélation négative implique un désaccord, en présence de variables externes

Comme CLV, selon le critère E, peut être adapté pour tenir compte de variables externes (§ 2.2), l'approche peut être également étendue dans le cas de l'optimisation du critère Q. Considérons la minimisation du critère :

$$Q^* = \frac{1}{n}\sum_{k=1}^{K}\sum_{j=1}^{p}\delta_{kj}\ \|\mathbf{x_j} - \mathbf{c_k}\|^{\mathbf{2}} \text{ sous la contrainte } \mathbf{c_k} = \mathbf{Z}\mathbf{b_k} \qquad (10)$$

$\mathbf{Z}$ désignant la matrice des variables externes centrées et éventuellement réduites.

Si K, le nombre de groupes, est fixé, dans chaque groupe G_k, la composante latente $\mathbf{c_k}$ recherchée est définie par :

$$\mathbf{c_k} = \mathbf{Z}\mathbf{b_k} \text{ avec } \mathbf{b_k} = (\mathbf{Z^t Z})^{-1}\,\mathbf{Z^t}\bar{\mathbf{x}}_\mathbf{k}$$

Ainsi, $\mathbf{c_k}$ est obtenu par régression linéaire multiple de $\bar{\mathbf{x}}_\mathbf{k}$ sur $\mathbf{Z}$, et pour une partition donnée en K_i groupes, la valeur minimale de Q^* vaut :

$$\begin{aligned} Q_i^* &= \frac{1}{n}\sum_{k=1}^{K_i}\sum_{j=1}^{p}\delta_{kj}\left\|\mathbf{x_j}-\mathbf{c_k}\right\|^2 \\ &= \frac{1}{n}\sum_{j=1}^{p}\mathbf{x_j^t x_j} - \frac{2}{n}\sum_{k=1}^{K_i}\sum_{j=1}^{p}\delta_{kj}\,\mathbf{x_j^t c_k} + \frac{1}{n}\sum_{k=1}^{K_i} p_k\,\mathbf{c_k^t c_k} \\ &= \frac{1}{n}\sum_{j=1}^{p}\mathbf{x_j^t x_j} - \frac{2}{n}\sum_{k=1}^{K_i} p_k\,\bar{\mathbf{x}}_\mathbf{k}^\mathbf{t}\mathbf{Z}\mathbf{b_k} + \frac{1}{n}\sum_{k=1}^{K_i} p_k\,\bar{\mathbf{x}}_\mathbf{k}^\mathbf{t}\mathbf{Z}(\mathbf{Z^tZ})^{-1}\mathbf{Z^tZ}\mathbf{b_k} \end{aligned}$$

ou encore :

$$Q_i^* = \sum_{j=1}^{p} Var(\mathbf{x_j}) - W_i^* \text{ où } W_i^* = \sum_{k=1}^{K_i} p_k\, Cov(\bar{\mathbf{x}}_\mathbf{k}, \mathbf{Zb_k}) \qquad (11)$$

En remarquant que : $Cov(\bar{\mathbf{x}}_\mathbf{k}, \mathbf{Zb_k}) = \frac{1}{n}\bar{\mathbf{x}}_\mathbf{k}^\mathbf{t}\mathbf{Z}(\mathbf{Z^tZ})^{-1}\mathbf{Z^t}\bar{\mathbf{x}}_\mathbf{k} = \frac{1}{n}\bar{\mathbf{x}}_\mathbf{k}^\mathbf{t}\mathbf{Z}(\mathbf{Z^tZ})^{-1}\mathbf{Z^tZ}(\mathbf{Z^tZ})^{-1}\mathbf{Z^t}\bar{\mathbf{x}}_\mathbf{k} = \frac{1}{n}\left(\mathbf{Z}(\mathbf{Z^tZ})^{-1}\mathbf{Z^t}\bar{\mathbf{x}}_\mathbf{k}\right)^\mathbf{t}\left(\mathbf{Z}(\mathbf{Z^tZ})^{-1}\mathbf{Z^t}\bar{\mathbf{x}}_\mathbf{k}\right)$, on peut également écrire :

$$W_i^* = \sum_{k=1}^{K_i} p_k\, Var(\mathcal{P}_\mathbf{Z}\bar{\mathbf{x}}_\mathbf{k})\,, \qquad (12)$$

$\mathcal{P}_Z = \mathbf{Z}(\mathbf{Z^tZ})^{-1}\mathbf{Z^t}$ désignant le projecteur sur l'espace engendré par $\mathbf{Z}$.

Il est cependant bien connu que le modèle de régression linéaire multiple n'est pas bien adapté aux situations dans lesquelles les prédicteurs (ici les variables $\mathbf{Z}$) sont fortement liés entre eux. Dans cette situation, la classification de variables basée sur le critère Q^* peut conduire à une partition non pertinente. Un critère alternatif (Vigneau et Qannari, 2003a) est alors à considérer. Ce critère alternatif, $\tilde{S}$, est similaire au critère généralisé $\tilde{T}$ (eq.(6)) mais dépend des covariances entre chacune des variables $\mathbf{x_j}$ et la variable latente du groupe auquel elle appartient, et non de leur carré :

$$\tilde{S} = \sum_{k=1}^{K}\sum_{j=1}^{p}\delta_{kj}\, Cov(\mathbf{x_j}, \mathbf{c_k}) \text{ sous les contraintes } \mathbf{c_k} = \mathbf{Za_k} \text{ et } \mathbf{a_k^t a_k} = 1 \qquad (13)$$

Pour ce nouveau problème d'optimisation, les composantes latentes des groupes $\mathbf{c_k}$ sont définies par :

$$\mathbf{c_k} = \mathbf{Za_k} \text{ avec } \mathbf{a_k} = \frac{\mathbf{Z}^\mathrm{t}\bar{\mathbf{x}}_\mathbf{k}}{\sqrt{\bar{\mathbf{x}}_\mathbf{k}^\mathrm{t}\mathbf{Z}\mathbf{Z}^\mathrm{t}\bar{\mathbf{x}}_\mathbf{k}}}$$

Dans un groupe G_k donné, la composante latente $\mathbf{c_k}$ est donc la première composante de la régression PLS1 de la variable centroïde du groupe $\bar{\mathbf{x}}_\mathbf{k}$ sur $\mathbf{Z}$.

D'un point de vue pratique, dans les deux cas (critère Q^* ou $\tilde{S}$), la structure des algorithmes hiérarchique et de partitionnement mis en œuvre sont identiques à ceux utilisés dans les autres situations.

2.5. Synthèse des différents cas inclus dans la méthode CLV

La méthode CLV intègre différentes options pour la classification de variables. On distingue le type de méthode (METHOD) qui dépend de la signification que l'on souhaite donner aux corrélations négatives entre variables : accord ou désaccord. La prise en compte de variables externes est gérée par le paramètre VARZ. Un dernier paramètre booléen (COLI) permet à l'utilisateur de tenir compte d'un éventuel problème de colinéarité ou quasi-colinéarité entre les variables externes. La synthèse des différents cas de figure est présentée dans le tableau 1.

TABLEAU 1

Les différents cas de figure de la méthode CLV, en fonction de trois paramètres : METHOD, VARZ et COLI ($\mathbf{c_k}$ désigne la composante latente dans le groupe G_k, $\mathbf{X_k}$, le tableau formé par les variables de ce groupe et $\bar{\mathbf{x}}_\mathbf{k}$, sa variable centroïde)

		METHOD=1 une forte corrélation positive ou négative signifie « accord »	**METHOD=2** une forte corrélation négative signifie « désaccord »
VARZ absent pas de variables externes		*critère : E* (eq.1) ou T (eq.2) (équivalence) $\mathbf{c_k}$: colinéaire à la première composante principale de $\mathbf{X_k}$	*critère : Q* (eq.8) $\mathbf{c_k}$: variable centroïde (ou moyenne) du groupe G_k
VARZ défini prise en compte de variables externes	**COLI=0** si pas de problème de colinéarité entre les variables externes	*critère : E^** (eq.4) ou T^* (eq.5) (équivalence) $\mathbf{c_k}$: première composante de l'ACPVI de $\mathbf{X_k}$ sur $\mathbf{Z}$.	*critère : Q^** (eq.10) $\mathbf{c_k}$: variable prédite par la régression de $\bar{\mathbf{x}}_\mathbf{k}$ sur $\mathbf{Z}$.
	COLI=1 si problème de colinéarité entre les variables externes	*critère : $\tilde{T}$* (eq.7) $\mathbf{c_k}$: première composante de la régression PLS2 de $\mathbf{X_k}$ sur $\mathbf{Z}$	*critère : $\tilde{S}$* (eq.13) $\mathbf{c_k}$: première composante de la régression PLS1 de $\bar{\mathbf{x}}_\mathbf{k}$ sur $\mathbf{Z}$

3. Applications

3.1. Tests psychotechniques de Holzinger

Le premier exemple considéré pour illustrer la classification de variables est basé sur un tableau de données relativement classique, utilisé notamment par Harman (1976). Il s'agit des résultats de 24 tests psychotechniques soumis à 145 enfants par Holzinger et Swineford en 1939. Le fichier des données a été téléchargé à partir du site du logiciel Mplus. Les variables correspondent à différents tests psychotechniques (tableau 2).

TABLEAU 2
Liste des 24 tests psychologiques et partitions

			CLV			**Varclus**		
num.	libellé	dénomination	G1	G2	G3	G1'	G2'	G3'
1	visual	perception visuelle	x			x		
2	cubes	cubes	x				x	
3	paper	papier	x			x		
4	flags	drapeaux	x				x	
5	general	information générale		x			x	
6	paragrap	compréhension, paragraphe		x			x	
7	sentence	compléter une phrase		x			x	
8	wordc	classification de mots		x			x	
9	wordm	compréhension de mots		x			x	
10	addition	addition			x			x
11	code	code			x			x
12	counting	comptage de points			x			x
13	straight	lettres majuscules			x			x
14	wordr	reconnaissance de mots	x			x		
15	numberr	reconnaissance de nombres	x			x		
16	figurer	reconnaissance de figures	x			x		
17	objectn	objet-nombre			x	x		
18	numberf	nombre-figure			x	x		
19	figurew	figure-mot			x	x		
20	deduct	déduction	x				x	
21	numeric	puzzles numériques			x			x
22	problemr	raisonnement	x				x	
23	series	compléter une série	x				x	
24	arithmet	problème d'arithmétique			x			x
		effectifs	*10*	*5*	*9*	*8*	*10*	*6*
		% inertie totale expliqué ...	*16.0*	*14.8*	*16.5*	*21.0*	*13.5*	*12.6*
		% inertie du groupe expliqué ...	*38.5*	*71.2*	*44.0*	*50.3*	*54.0*	*37.9*
		par la variable latente du groupe						

Comme l'intérêt porte sur la corrélation entre les tests, les données ont été standardisées. On considère ici qu'une corrélation positive ou négative correspond à un accord. La procédure Varclus de SAS a également été mise en œuvre à des fins de comparaison. La figure 2 montre l'évolution du critère T (eq.3) ramené à l'inertie totale des données (il s'agit par conséquent du pourcentage d'inertie totale expliqué par les variables CLV) en fonction du nombre de groupes. Ces résultats donnés pour les méthodes CLV et Varclus révèlent une grande similitude de performance des deux approches.

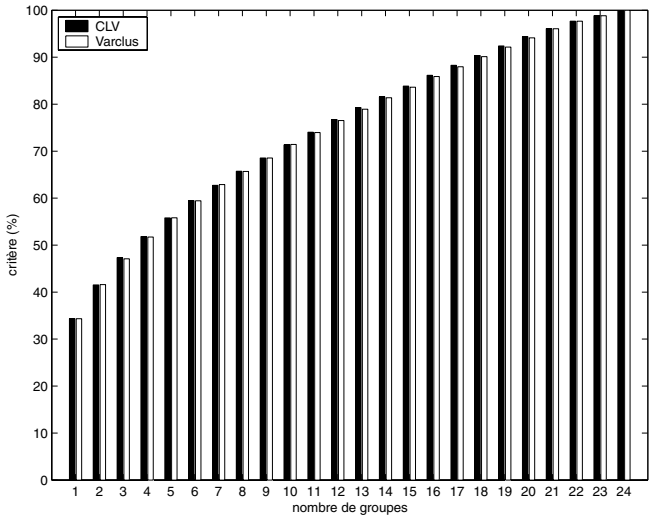

FIGURE 2

Évolution du pourcentage d'inertie expliquée par les variables latentes de groupes, en fonction du nombre de groupes pour les deux méthodes de classification : méthode CLV et méthode Varclus de SAS

Cependant la similitude de la valeur du critère d'évaluation des partitions ne signifie pas que les partitions obtenues pour un nombre K de groupes fixé soient très ressemblantes. Pour illustrer ceci, considérons une partition en K=3 groupes. Ce choix s'appuie sur l'observation du dendrogramme de la phase de classification hiérarchique des variables de l'approche CLV (figure 3). L'algorithme de partitionnement de la méthode CLV n'a pas modifié la partition en trois groupes obtenue par coupure du dendrogramme. Celle-ci est décrite dans le tableau 2. Les résultats obtenus avec la procédure Varclus figurent également dans ce tableau. Les deux partitions permettent d'expliquer respectivement 47.3 % et 47.1 % de l'inertie totale mais correspondent à des regroupements de variables assez différents. Si l'on décompose le critère global par groupe, c'est-à-dire si on évalue le pourcentage *d'inertie totale* expliquée par chacune des variables latentes de groupe (avant dernière ligne du tableau 2) on remarque que l'importance relative de chaque groupe est plus homogène lorsque la méthode CLV est mise en œuvre que lorsque la méthode Varclus est utilisée. Le constat change si on s'intéresse plutôt au pourcentage *d'inertie du groupe* restitué par la variable latente de ce groupe (dernière ligne du tableau 2). Cette différence

est certainement due aux différences de principe des deux approches, le critère de formation des groupes considéré par CLV étant fondé sur la valeur propre $\lambda_1^{(k)}$ de la première composante principale de chaque groupe G_k alors que la procédure Varclus utilise une démarche basée sur le pourcentage d'inertie du groupe associé à cette première composante principale, à savoir $\lambda_1^{(k)}/p_k$ (lorsque les variables sont standardisées). Par ailleurs, il est à noter que les conclusions tirées par Harman (1976) se recoupent davantage avec les résultats obtenus par la méthode CLV qu'avec ceux obtenus avec la méthode Varclus.

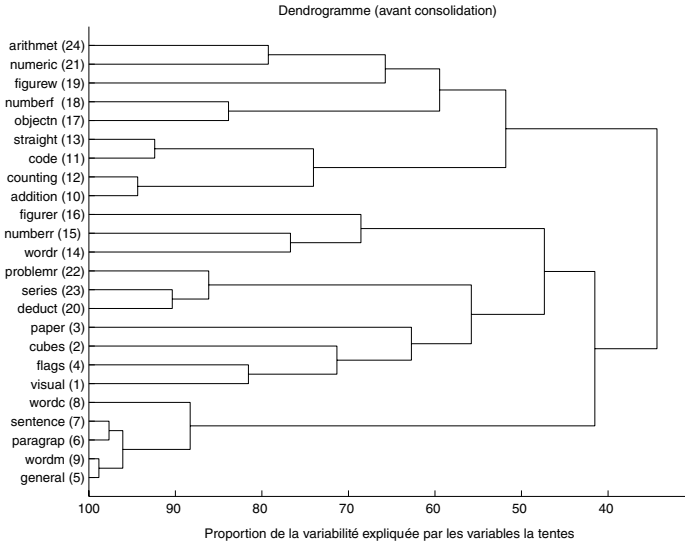

FIGURE 3
Dendrogramme de la classification ascendante hiérarchique des variables avec l'approche CLV (cas où le signe de la corrélation n'est pas pertinent). Exemple des tests psychologiques

3.2. Classification de consommateurs en fonction de leurs préférences, avec prise en compte de variables sensorielles

Dans ce second exemple, l'objectif est d'identifier des groupes dans un panel de consommateurs auxquels il a été demandé de donner une note de préférence à un ensemble de produits. Si en plus des données de préférence, on dispose d'une description sensorielle des produits, l'intérêt de la démarche dite de cartographie externe des préférences, est d'exhiber des groupes de consommateurs ayant des préférences similaires et de pouvoir comprendre les déterminants sensoriels de ces préférences. La méthode de classification des variables CLV a été, au départ, développée afin de répondre à ce type de problématique (Vigneau *et al.*, 2001 ; Vigneau et Qannari, 2002).

Nous allons considérer une partie des données collectées dans le cadre d'un projet européen (ESN : European Sensory Network, 1996) portant sur l'analyse

sensorielle de cafés réalisée simultanément dans plusieurs pays d'Europe. Huit variétés de café ont été évaluées par des consommateurs de toute l'Europe. Pour simplifier l'illustration de la méthode de classification de variables, nous avons considéré seulement les résultats de $p = 233$ consommateurs français, allemands et polonais. Pour chacun des cafés, on dispose également de son profil sensoriel sur la base de $q = 18$ descripteurs (voir tableau 3). Ces descripteurs sensoriels forment la matrice de données externes, **Z**.

TABLEAU 3
Liste des 18 descripteurs sensoriels

Odeur	*Goût*	*En bouche*	*Arrière-goût*
o-intensité	g-intensité	b-fin-épais	arrg-intensité
o-chocolat	g-épicé		
o-verte	g-brûlé		
o-grillée	g-aigre		
o-moisie	g-chocolat		
o-sucrée	g-metal		
o-parfumée	g-amer		
o-caramel	g-sucré		

L'objectif de l'étude est de séparer l'ensemble des consommateurs en groupes au sein desquels les personnes ont des préférences voisines. Pour cela, la méthode CLV a été appliquée en considérant qu'une corrélation négative indique un désaccord. La prise en compte des données externes est réalisée sur la base du critère $\tilde{S}$ (eq.13).

Le dendrogramme obtenu avec la méthode CLV (figure 4) montre clairement une partition en deux groupes. Avant consolidation à l'aide de l'algorithme de partitionnement, ces deux groupes comportaient 83 et 150 consommateurs et la valeur du critère d'évaluation de la partition $\tilde{S}$ (eq.13) valait 833.76. Après consolidation, le premier de ces groupes (G1) rassemble 82 consommateurs (35 %), le second G2, 151 consommateurs (65 %) et la valeur du critère d'évaluation de la partition atteint la valeur de 843.02.

La description des deux groupes est donnée dans la figure 5. Les variables latentes associées respectivement aux deux groupes sont représentées dans la partie droite de cette figure. Il est clair que ces deux variables latentes décrivent deux directions de préférence opposées. Les différences d'appréciation s'expliquent bien en fonction des descripteurs sensoriels (voir la partie gauche de la figure 5). Les caractéristiques appréciées par les consommateurs du groupe G1 sont l'intensité de l'odeur, du goût et de l'arrière-goût, avec des goûts prononcés (brûlé, épicé, amer) et une texture épaisse. Les consommateurs du groupe G2 valorisent les aspects sucrés et chocolatés des cafés ainsi que l'odeur caramel. La composition de ces deux groupes de consommateurs en fonction de leur nationalité ne permet pas de relier clairement les différences de goût et le pays d'origine, même si en majorité les consommateurs qui préfèrent les produits corsés (G1) sont d'origine polonaise.

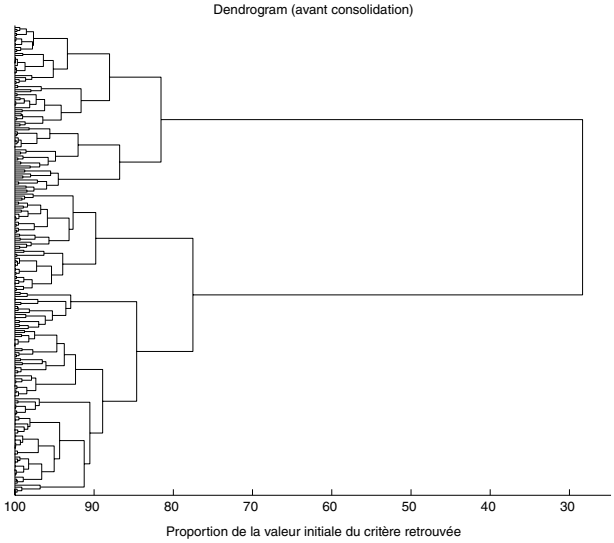

FIGURE 4
Dendrogramme de la classification ascendante hiérarchique des variables avec l'approche CLV (cas où une corrélation négative indique un désaccord et avec variables externes). Exemple des cafés

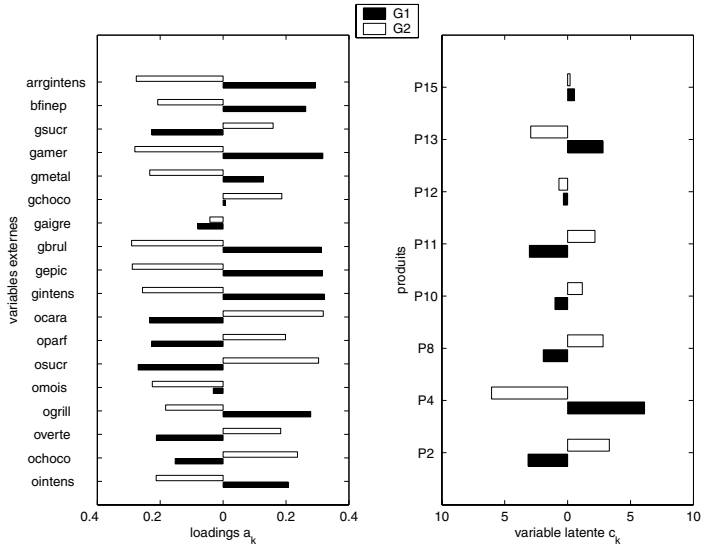

FIGURE 5
A gauche : profils des variables latentes des groupes G1 et G2 en fonction des descripteurs sensoriels.
A droite : coordonnées des produits selon ces variables latentes

4. Conclusion

Avec l'émergence de tableaux de données de plus en plus grands et l'acquisition automatique de nombreux caractères pour décrire un système, la classification de variables est une préoccupation d'actualité dans de nombreux domaines d'application. L'approche CLV que nous proposons dans ce contexte intègre plusieurs options qui permettent de la mettre en œuvre dans différentes situations. On distingue ainsi la classification locale de variables et la classification de variables autour d'axes.

Concrètement, les objectifs recherchés sont voisins de ceux de la procédure Varclus bien connue des utilisateurs du logiciel SAS. La démarche est cependant plus simple, ce qui permet notamment d'implémenter la méthode CLV sans difficulté majeure dans différents environnements (Une macro écrite sous SAS ainsi qu'une fonction en langage Matlab sont disponibles auprès des auteurs). Les critères optimisés dans le cadre de CLV sont par ailleurs clairement définis. De ce fait, une extension des démarches pour la réalisation d'une classification de variables avec prise en compte de leurs relations linéaires avec d'autres variables a été proposée. Dans ce cas, la structure de la matrice de variance-covariance des variables externes ayant une incidence en terme d'interprétabilité des partitions obtenues, des alternatives sont envisagées en cas de forte colinéarité entre variables externes.

Lorsque les groupes de variables sont organisés selon un axe n'ayant pas d'orientation spécifique, la démarche CLV permet de compléter les résultats de l'ACP. La contrainte d'orthogonalité des variables latentes est relaxée au profit d'une meilleure association entre les variables observées et les variables latentes et d'une plus grande facilité d'interprétation. Si des variables externes sont prises en compte, cette complémentarité existe vis-à-vis de la Régression PLS2 ou l'ACP sur Variables Instrumentales. Dans le cas où l'orientation des variables latentes de groupes a un sens concret, et avec intégration de variables externes, la démarche CLV s'apparente à une agrégation de variables autour de modèles qui sont des modèles de régression linéaire multiple ou de régression PLS1.

Remerciements

Les auteurs tiennent à remercier P. L. Gonzalez et G. Saporta pour l'intérêt particulier qu'ils ont porté à ce travail.

Références

ABDALLAH H., SAPORTA G. (1998), Classification d'un ensemble de variables qualitatives, *Revue de Statistique Appliquée*, XLVI(4), 5-26.

AL-KANDARI N. M., JOLLIFFE I. T. (2001), Variable selection and interpretation of covariance principal components, *Communications in Statistics : Simulation and Computation*, 30, 339-354.

DERQUENNE C. (1997), Classification de variables qualitatives, XXIX^{e} Journées de l'ASU, Carcasonne.

DIDAY E. (1971), Une nouvelle méthode en classification automatique et reconnaissance des formes : La méthode des nuées dynamiques, *Revue de Statistique Appliquée*, XIX(2), 19-33.

DIDAY E. (1974), Introduction à l'analyse factorielle typologique, *Revue de Statistique Appliquée*, XXII(4), 29-38.

DIDAY E. (1976), Sélection typologique de paramètres, Rapport de Recherche n° 188, INRIA, Le Chesnay, France.

ESN (1996), A European Sensory and Consumer Study : A Case Study on Coffee, ESN, Gloucestershire, UK.

GREENHOFF K., MACFIE H. J. H. (1994), Preference mapping in practice, Measurement of food preferences. H. J. H. Macfie and D. M. H. Thomson, Eds, Blackie academic & professional : London, 137-166.

GUO Q., WU W., MASSART D. L., BOUCON C., DE JONG S. (2002), Feature selection in principal component analysis of analytical data, *Chemometrics and Intelligent Laboratory Systems*, 61, 123-132.

HARMAN H. H. (1976), Factor Analysis, Third edition, The University of Chicago Press, Chicago and London.

JOLLIFFE I. T. (1972), Discarding variables in a principal component analysis. I : Artificial data, *Applied Statistics*, 21, 160-173.

KRZANOWSKI W. J. (1987), Selection of variables to preserve multivariate data structure, using principal components, *Applied Statistics*, 36, 22-33.

MCCABE G.P. (1984), Principal variables, *Technometrics*, 26, 137-144.

QANNARI E. M., VIGNEAU E., COURCOUX P. (1998), Une nouvelle distance entre variables; application en classification, *Revue de Statistique Appliquée*, XLVI(2), 21-32.

RAO C. R. (1964), The use and the interpretation of principal component analysis in applied research, *Sankhya*, A, 26, 329-358.

SABATIER R. (1987), Méthodes factorielles en analyse des données, approximations et prise en compte des variables concommitantes, Thèse d'état, USTL, Montpellier.

SAS/STAT (1990), User's guide, Version 6, Vol.2, SAS Institute Inc. : Cary, North Carolina.

SOFFRITTI G. (1999), Hierarchical clustering of variables : a comparison among strategies of analysis, *Communications in Statistics : Simulation and Computation*, 28, 977-999.

VIGNEAU E., QANNARI E.M., PUNTER P. H., KNOOPS S. (2001), Segmentation of a panel of consumers using clustering of variables around latent directions of preference, *Food Quality and Preference*, 12, 359-363.

VIGNEAU E., QANNARI E.M.(2002), Segmentation of consumers taking account of external data. A clustering of variables approach, *Food Quality and Preference*, 13, 515-521.

VIGNEAU E., QANNARI E.M. (2003a), Clustering of variables around latent components, *Communications in Statistics : Simulation and Computation*, 32 (4), 1131-1150.

VIGNEAU E., QANNARI E.M. (2003b), Clustering of variables around latent components and PLS regression : application to sensory data, PLS'03 International Symposium, M. Vilares, M. Tenenhaus, P. Coelho, V. Esposito Vinzi, A. Morineau (eds.), DECISIA.