REVUE DE STATISTIQUE APPLIQUÉE

B. ESCOFIER
J. PAGES

Méthode pour l'analyse de plusieurs groupes de variables. Application à la caractérisation de vins rouges du Val de Loire

Revue de statistique appliquée, tome 31, n° 2 (1983), p. 43-59
<http://www.numdam.org/item?id=RSA_1983__31_2_43_0>

METHODE POUR L'ANALYSE DE PLUSIEURS GROUPES DE VARIABLES – APPLICATION A LA CARACTERISATION DE VINS ROUGES DU VAL DE LOIRE

B. ESCOFIER (*) J. PAGES (**)

1. INTRODUCTION

Nous proposons ici une méthode d'analyse de tableaux de données croisant un ensemble d'individus et plusieurs groupes de variables. Il est très courant d'étudier ce type de tableaux qui peuvent comprendre des variables numériques, ou des variables qualitatives. L'unité d'un groupe provient de ce que les variables qui le constituent se réfèrent à un même thème, à une même date. . .

Dans l'exemple qui nous guidera au cours de cette présentation, les individus sont 19 vins proposés à des dégustateurs professionnels. Ces vins sont jugés selon 21 critères qui se rapportent aux trois sens de la vue, de l'odorat et du goût. Le nombre de critères pour chacun de ces trois sens est respectivement de 5, 9 et 7. Le tableau étudié prend en compte les moyennes des "notes" attribuées à ces vins par les 24 dégustateurs.

Les questions que l'on peut se poser devant ce type de tableau sont diverses et ont donné lieu à des méthodes très différentes. Citons par exemple l'analyse multicanonique (Carroll), la méthode STATIS (Escoufier) [4 et 6].

Nous abordons l'ensemble de ces questions, et proposons une solution optimale (en un certain sens) à chaque problème. Nous aboutissons à une méthode unique, basée sur une anlayse en composante principale du tableau de données. Dans cette analyse, pour équilibrer le rôle des groupes de variables, chacun d'entre eux est pondéré. Les résultats obtenus permettent en particulier une représentation graphique cohérente :

– du nuage d'individus (ici les vins) défini par l'ensemble des variables ;
– de nuages d'individus définis par chaque groupe de variables (ici 3 nuages) ;
– du nuage des variables ;
– de points (ici 3) représentant chaque groupe de variables.

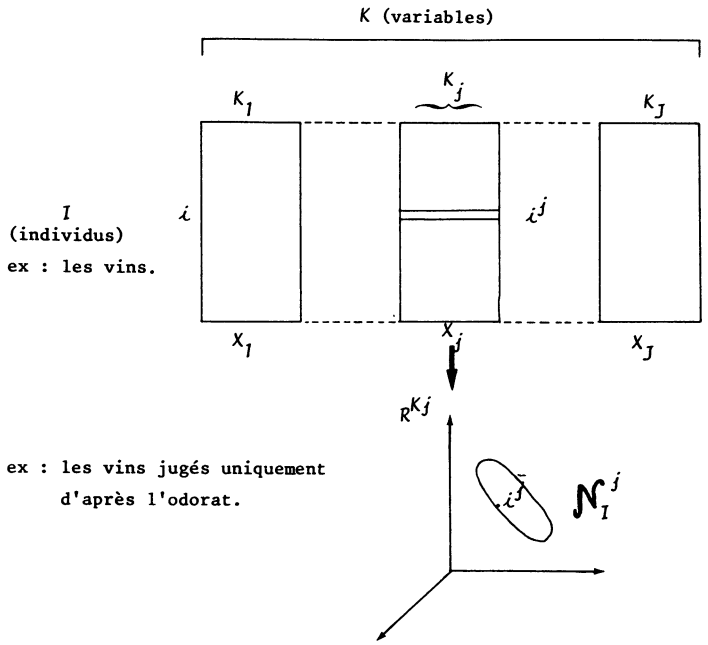
L'exposé complet de la méthode se trouve dans [3]. Nous nous contentons ici d'en exposer les principes et de commenter un exemple d'application.

2. NOTATIONS

Le tableau de données X croise un ensemble d'individus noté I, et un ensemble de variables noté K. Ce dernier est divisé en J groupes notés K_j, définissant ainsi J sous tableaux notés X_j.

(*) IRISA, Campus de Beaulieu, avenue du Général Leclerc, 35042 Rennes Cédex.
(**) ENSAR, 69 route de Saint Brieuc, 35042 Cédex.

Mots-Clés : **Analyse en composantes principales, Analyse des correspondances multiples, Analyse multicanonique, Comparaison de tableaux.**



A chaque groupe de variables K_j, correspond un nuage d'individus situé dans l'espace R^{Kj}. On note $\mathcal{N}_I^j$ ce nuage et i^j le point représentant i dans ce nuage.

Des poids peuvent être affectés aux individus et aux variables. Nous notons respectivement D et M les matrices diagonales des poids p_i des individus et des poids m_k des variables et M_j la matrice diagonale des poids des variables du groupe K_j.

Nous notons W_j le produit $X_jM_jX_j'$ où X_j' est la transposée de X_j. C'est la matrice d'inertie du nuage des variables du groupe K_j. Nous notons de même $W = XMX'$.

3. COMPARAISON DES NUAGES D'INDIVIDUS A L'AIDE D'UNE REPRESENTATION SIMULTANEE

Dans l'exemple, nous sommes en présence de 3 nuages représentant les 19 vins jugés selon des critères visuels, olfactifs et gustatifs. Un des problèmes posés dans l'étude de ce type de tableau est la comparaison de ces nuages. Plus précisément, on souhaite savoir si, généralement, deux vins proches du point de vue de l'oeil le sont aussi du point de vue du nez et du goût. Si cette situation est réalisée, il sera intéressant de mettre en évidence des exceptions. Une autre façon d'aborder ce problème consiste à rechercher des dimensions communes à ces trois nuages (i.e. des "facteurs communs").

Une représentation simultanée de ces trois nuages sur des espaces de petite dimension permet de répondre en partie à ces questions, à condition que cette

représentation possède certaines qualités. Plus précisément il est possible de dégager 4 propriétés essentielles pour qu'une telle représentation simultanée soit utilisable :

P(1) La représentation de chaque nuage d'individus doit être une projection de ce nuage.

P(2) La qualité de représentation de chacun de ces nuages doit être assez bonne ; cette qualité est mesurée par le pourcentage d'inertie extrait par la projection.

P(3) Les points représentant le même individu dans chacun des nuages doivent être assez proches. Ceci permet de dégager des facteurs communs, s'il en existe, et de faciliter la comparaison des positions des individus dans les différents nuages.

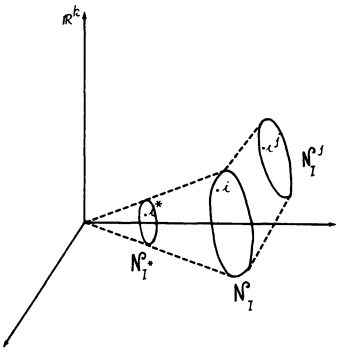
Il est évident que les deux dernières propriétés ne peuvent généralement pas être optimisées simultanément et qu'il faudra trouver un compromis entre elles. En effet, la meilleure qualité de représentation de chacun des nuages sur un plan est donnée par les deux premiers facteurs de l'analyse des variables de son groupe. La superposition de ces plans ne permet pas de comparer facilement les nuages : une ressemblance peut être masquée par une rotation ou une inversion dans l'ordre des facteurs. A l'opposé, une représentation où tous les points seraient confondus optimiserait la propriété P(3) mais n'aurait aucun intérêt.

P(4) Pour faciliter la comparaison des nuages, surtout s'ils sont assez nombreux, il est nécessaire qu'apparaisse aussi la projection d'un nuage "moyen" compromis entre tous les nuages. Pour que ce nuage joue bien son rôle de compromis il faut que chacun de ses points soit projeté au centre de gravité des points homologues de tous les nuages.

Solution proposée

Considérons l'espace $\mathbf{R}^K$ engendré par l'ensemble de toutes les variables. Cet espace, muni de la métrique diagonale des poids des variables, est la somme directe de J sous-espaces, isomorphes aux espaces $\mathbf{R}^{Kj}$, et orthogonaux entre eux. Il est ainsi possible de plonger les J nuages $\mathcal{N}_I^j$, situés dans les $\mathbf{R}^{Kj}$, dans le même espace $\mathbf{R}^K$. Cette représentation est artificielle, puisque ces nuages évoluent dans des sous-espaces orthogonaux entre eux, mais servira de base à leur représentation simultanée.

Dans cet espace $\mathbf{R}^K$, nous pouvons construire un nuage moyen noté $\mathcal{N}_I^*$ compromis entre les J nuages. Le point i^* représentant un individu i dans ce nuage moyen est situé au barycentre des J points i^j représentant i dans les nuages $\mathcal{N}_I^j$.



Pour obtenir une représentation simultanée des $\mathcal{N}_I^j$ dans un espace de petite dimension, nous allons les projeter sur un sous-espace de $\mathbf{R}^K$. La propriété P(1) sera ainsi automatiquement vérifiée et la projection du nuage moyen sur ce même sous espace assure la réalisation de la propriété P(4). Ce principe étant acquis, seules les propriétés P(2) et P(3) induisent le choix du sous-espace de $\mathbf{R}^K$. Or, ces propriétés peuvent s'exprimer en termes d'inertie. Une bonne qualité de représentation des $\mathcal{N}_I^j$ (propriété P(2)) implique que l'inertie de leur projection soit grande ; il faut donc que l'inertie de l'union de ces nuages soit grande. L'inertie de ce gros nuage comprenant I x J points peut se décomposer selon le théorème de Huyghens, en une inertie inter et une inertie intra. L'inertie inter est l'inertie des centres de gravité des J points i^j représentant le même individu, soit l'inertie du nuage moyen $\mathcal{N}_I^*$. L'inertie intra est l'inertie de ces points i^j autour de leur centre de gravité i^*. Pour que ces points soient proches entre eux (propriété P(3)), il faut rendre cette inertie "intra" faible.

Pour rendre à la fois l'inertie totale grande et l'inertie intra faible nous choisissons les sous-espaces rendant maximum l'inertie inter. Puisque cette inertie inter est celle du nuage $\mathcal{N}_I^*$, nous projeterons les $\mathcal{N}_I^j$ et $\mathcal{N}_I^*$ sur les sous espaces engendrés par les premiers axes d'inertie de $\mathcal{N}_I^*$.

La solution numérique est très simple. En effet, $\mathcal{N}_I^*$ se confond, au facteur d'homothétie 1/J près, avec le nuage associé à l'ensemble de toutes les variables. L'analyse revient à réaliser l'analyse en composantes principales (s'il s'agit de variables numériques) du tableau X et à projeter en éléments supplémentaires les individus i^j.

Pour être réellement utilisable, cette représentation simultanée doit être complétée par des indices mesurant les qualités des représentations des nuages et leur ressemblance. Nous en présenterons quelques uns dans le commentaire de l'exemple.

Remarque 1

La projection du nuage moyen sur un axe u s'écrit, au facteur 1/J près, $\mathcal{F} = XMu$. Celle d'un nuage $\mathcal{N}_I^j$ s'écrit $\mathcal{F}^j = X_jMu$ où X_j est un tableau de dimension I x K constitué par X_j complété par des zéros.

Si u est un axe d'inertie, la dualité de l'analyse en composantes principales permet d'écrire le vecteur normé u en fonction de $\mathcal{F}$

$$u = (1/\lambda)\, X'D\, \mathcal{F}$$

où X' est le transposé de X et λ la valeur propre associée à $\mathcal{F}$ et u.

D'où :

$$\begin{aligned} \mathcal{F}^j &= (1/\lambda)\, \tilde{X}_j M X' D \mathcal{F} \\ &= (1/\lambda)\, W_j D \mathcal{F} \end{aligned}$$

Ainsi, la projection du nuage $\mathcal{N}_I^j$ dans $\mathbf{R}^K$ s'interprète dans l'espace $\mathbf{R}^I$ associé au même tableau : à un coefficient près, on applique l'opérateur W_jD aux composantes principales.

Remarque 2

Si u_j est la projection de u sur le sous espace $\mathbf{R}^{Kj}$ et θ_j l'angle entre u et u_j alors la projection $\mathcal{F}^j$ de $\mathcal{N}_I^j$ sur u est égale, au coefficient $\cos\theta_j$ près, à sa projection sur u_j.

Bien que la qualité de représentation de $\mathcal{N}_I^j$ soit meilleure sur u_j que sur u, c'est la projection sur u qui est utilisée dans la représentation simultanée. En effet : les projections sur $\mathbf{R}^{Kj}$ de deux axes orthogonaux u et u' ne sont généralement pas orthogonales ; d'autre part, les projections sur les u_j ne vérifient pas la propriété P(3) et n'optimisent pas P(4).

Remarque 3

Dans le cas de groupes réduits à une seule variable, la méthode proposée dans ce paragraphe est identique à une analyse en composantes principales normée.

4. COMPARAISON DES NUAGES DE VARIABLES

Le paragraphe précédent nous a conduit à une analyse en composantes principales de l'ensemble des variables. Cette analyse inclut une représentation des variables qui permet d'étudier les corrélations entre variables de même groupe mais aussi entre variables de groupes différents. On effectue ainsi une comparaison point par point des nuages de variables.

Une autre approche, consiste à caractériser les groupes de variables par leurs composantes principales.

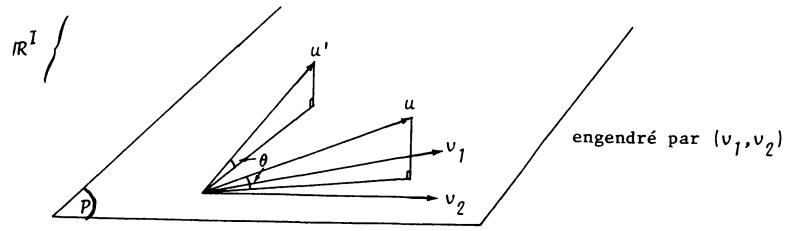
Une A.C.P. de ces composantes, pondérées par leur inertie, est un moyen efficace pour comparer les groupes de variables. Or, cette analyse est équivalente à celle de l'ensemble des variables. Il suffit donc d'introduire les composantes principales des groupes de variables en tant qu'élément supplémentaire de l'analyse de l'ensemble des variables.

5. ANALYSE DES LIAISONS ENTRE LES GROUPES DE VARIABLES

Nous adoptons maintenant un tout autre point de vue : celui de l'analyse canonique généralisée, dont le but est de rechercher des combinaisons linéaires des variables de chaque groupe les plus liées entre elles. La technique la plus classique, basée sur une idée de Carroll [1 et 5] passe par l'intermédiaire de variables générales liées à l'ensemble des groupes. La mesure de liaison entre une variable et un groupe est le coefficient de corrélation multiple. La première variable générale est la variable rendant maximum la somme des carrés des corrélations multiples ; on calcule ensuite les combinaisons linéaires des variables de chaque groupe les plus corrélées à cette variable générale ; on itère alors le procédé en ajoutant une contrainte d'orthogonalité sur les variables générales.

Le choix du coefficient de corrélation multiple comme mesure de liaison présente quelques inconvénients car il ne tient compte que du sous-espace de $\mathbf{R}^I$ engendré par les variables du groupe. Ainsi par exemple, les variables u et u' du graphique ci-dessous ont le même coefficient de corrélation multiple avec les variables (v_1, v_2) bien que u' soit très peu lié à ces variables.

D'autre part lorsque v_1 et v_2 sont très corrélées, le sous espace engendré, et par voie de conséquence le cœfficient de corrélation multiple avec une variable quelconque est très instable.



C'est pourquoi, nous proposons de suivre la démarche de Carroll en remplaçant le carré du cœfficient de corrélation multiple par une autre mesure pour apprécier la liaison entre une variable u et un ensemble de variables $\{v_k/k = 1, K_j\}$; cette mesure s'écrit :

$$\text{Liaison } [u,(v_1, v_2, \ldots v_{K_j})] = \sum_{k\in K_j} \text{inertie des projections des } v_k \text{ sur u}$$

$$= \sum_{k\in K_j} m_k \langle u, v_k \rangle^2 \qquad (1)$$

Cette mesure croît lorsque u, à angle égal avec le sous espace engendré par les v_k, s'approche d'une direction de grande inertie des v_k. Elle est beaucoup plus stable que la corrélation multiple.

La solution de cette analyse multicanonique particulière est très simple ; la variable qui maximise la somme des liaisons (telles que nous venons de les définir) avec tous les groupes est clairement la première composante principale de l'ensemble des variables. L'itération avec contrainte d'orthogonalité conduit aux composantes principales suivantes.

Nous aboutissons à la méthode proposée au paragraphe 3 qui se trouve ainsi enrichie d'une autre interprétation et complétée par un indice de liaison entre une composante et un groupe.

Les composantes principales sont les vecteurs propres de la matrice WD où W est la somme des J matrices d'inertie W_j. L'opérateur associé WD est donc la somme des opérateurs W_jD. Or, dans l'analyse multicanonique classique, les variables générales sont les vecteurs propres de la somme des opérateurs de projections orthogonales sur les sous espaces engendrés par les variables de chaque groupe. Le changement d'indice de liaison se traduit donc par le remplacement par les W_jD de ces opérateurs de projections orthogonales que nous notons P_j.

Dans la deuxième phase de l'analyse canonique généralisée au sens de Carroll, on applique aux variables générales les opérateurs de projection P_j ; on obtient ainsi les combinaisons linéaires de variables de chaque groupe les plus corrélées aux variables générales. Dans la méthode que nous proposons, nous choisissons, à la deuxième phase, d'appliquer aux variables générales les opérateurs W_jD. Ce choix se justifie par plusieurs raisons :

- déjà dans la première phase nous avons indiqué les rôles parallèles joués par les opérateurs P_j et W_jD dans la méthode de Carroll et celle que nous proposons.

- en appliquant l'opérateur W_jD plutôt que P_j, on obtient des variables qui représentent mieux les groupes.

En effet, on peut montrer que W_jDu correspond à une direction de plus grande inertie que P_ju. (sauf dans le cas extrême où P_ju est colinéaire à un vecteur propre de W_jD, auquel cas P_ju et W_jDu ont des directions identiques).

Rappelons que c'est un souci analogue qui nous a guidé dans le choix de la mesure de liaison utilisée dans la première phase.

• les variables obtenues s'interprètent, à un facteur d'homothétie près, comme les coordonnées des points du nuage $\mathcal{N}_I^j$ dans la représentation décrite au paragraphe 3. (le résultat y est démontré dans la remarque située à la fin de ce paragraphe).

Ceci permet d'associer à l'analyse des liaisons des groupes une représentation simultanée des nuages jouissant en elle même de bonnes propriétés et réciproquement.

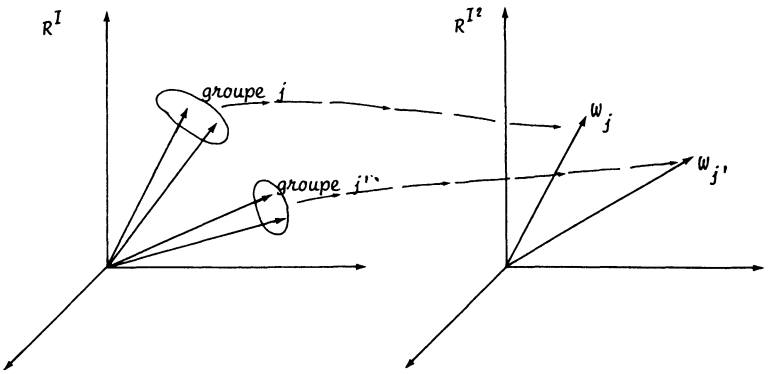
6. COMPARAISON GLOBALE DES GROUPES DE VARIABLES

Un point de vue encore différent réside dans la comparaison globale des groupes de variables : quels sont les groupes qui se ressemblent ? Certains groupes peuvent-ils être considérés comme intermédiaires entre deux autres ?

Cette comparaison globale nécessite la définition d'une mesure de distance ou de liaison entre groupes. Or, il existe une mesure bien adaptée qui peut d'ailleurs être introduite de multiples façons : par exemple, on choisira de représenter chaque groupe de variables K_j par la matrice W_j qui s'interprète, à la fois, comme la matrice des produits scalaires entre les points du nuage $\mathcal{N}_I^j$ et comme la matrice d'inertie du nuage des variables K_j dans l'espace $\mathbf{R}^I$. Les matrices W_j appartiennent à un espace de dimension I^2 que nous noterons $\mathbf{R}^{I2}$ et que l'on munit du produit scalaire induit par le produit scalaire D de $\mathbf{R}^I$. Soit :

$$\langle W_j, W_{j'} \rangle = \text{trace}\,(W_j D W_{j'} D)$$

Les matrices W_j auxquelles nous nous intéressons sont toujours positives ou semi définies positives, ce produit scalaire est toujours positif pour ce type de matrices.



Ce produit scalaire constitue une mesure de liaison entre les groupes de variables. Il suffit, pour s'en convaincre de considérer les deux cas particuliers suivants :

– si chacun des groupes ne comporte qu'une variable centrée réduite, il conduit au carré de leur coefficient de corrélation.

– si l'un des groupes est réduit à une variable centrée réduite, il conduit à la mesure de liaison utilisée au paragraphe précédent.

Dans le contexte d'analyse factorielle qui est le nôtre, afin de comparer facilement les points W_j, il est naturel de les projeter sur un sous-espace de petite dimension de $\mathbf{R}^{I2}$. Un ajustement au moindre carré des W_j, équivalent à une analyse en composantes principales, donne une bonne idée des distances entre les W_j mais aucun élément d'interprétation de ces distances.

Afin d'obtenir des éléments d'interprétation en termes de variables (ou d'individus) nous proposons de projeter les W_j sur un sous-espace engendré par des éléments de $\mathbf{R}^{I2}$ d'un type très particulier : des matrices associées à un groupe de variables composé d'un seul élément. Ainsi chaque vecteur de base du sous-espace (dans $\mathbf{R}^{I2}$) correspondra à une variable (dans $\mathbf{R}^{I}$) et les proximités des projections de W_j s'interprèteront en s'appuyant sur ces variables.

Nous chercherons donc un premier vecteur de $\mathbf{R}^{I2}$ noté Z_1 associé à un unique vecteur centré normé de $\mathbf{R}^{I}$ noté z_1 ($Z_1 = z_1 \, D \, z_1'$). Le critère d'ajustement au moindre carré conduirait à chercher un vecteur Z_1 rendant maximum la somme des carrés des projections des W_j associés aux J groupes de variables K_j. Nous avons préféré rendre maximum la somme de ces projections. Ceci est licite puisque, le produit scalaire entre W_j et Z_1 est positif et permet d'assurer la cohérence avec les autres points de vue. En effet, la projection de W_j sur Z_1 vaut :

$$\langle W_j , Z_1 \rangle = \sum_{k \in K_j} m_k \langle v_k , z_1 \rangle^2$$

Si Z_1 rend maximum la somme des projections des $W_j (1 \leqslant j \leqslant J)$, alors z_1 rend maximum la somme de l'inertie des projections des $v_k (k \in K)$. Le vecteur z_1 est donc la première composante principale de l'ensemble des variables et la coordonnées de W_j sur Z_1 est égale à la contribution du groupe de variables j à l'inertie de cette composante.

La recherche d'un vecteur Z_2 de $\mathbf{R}^{I2}$, orthogonal à Z_1, associé à un unique vecteur z_2 de $\mathbf{R}^{I}$, optimisant le même critère conduit de la même façon à la seconde composante principale etc. L'analyse proposée est donc parfaitement liée aux autres points de vue et le calcul des projections des W_j se déduit des résultats de l'A.C.P.

Les projections des W_j sur le sous-espace de $\mathbf{R}^{I2}$ défini ci-dessus, étant des combinaisons linéaires d'une même suite de vecteurs de $\mathbf{R}^{I2}$ associés chacun à une variable, sont des sommes pondérées de produits scalaires induit chacun par une seule variable.

Nous rejoignons ainsi un tout autre aspect de la comparaison de tableaux, celui qui est proposé dans le modèle INDSCAL [2] où l'on cherche à décomposer un ensemble de matrices de proximités (ou de produits scalaires) suivant des facteurs qui leurs sont communs.

La méthode proposée fournit une interprétation géométrique de la formalisation et de l'estimation des paramètres du modèle INDSCAL : les variables définissant le sous espace de $\mathbf{R}^{I2}$ sur lequel on projette les W_j constituent les facteurs communs ; les coordonnées des W_j sont les poids affectés par chacun des W_j à ces facteurs.

7. LA PONDERATION DES GROUPES DE VARIABLES

Quel que soit le point de vue adopté lors du traitement du type de tableau auquel nous intéressons, il se pose un problème essentiel non abordé jusqu'ici : la pondération des groupes de variables, indispensable pour équilibrer le rôle des différents groupes dans l'analyse (en particulier lorsque le nombre de variables varie beaucoup d'un groupe à l'autre).

Nous proposons de pondérer les variables d'un groupe par un même coefficient : l'inverse de la plus grande valeur propre de l'analyse de ce groupe. Cette pondération a pour effet de rendre égale à 1 l'inertie maximum dans une direction donnée des nuages K_j de variables (ou $\mathcal{N}_I^j$ d'individus).

Avec cette pondération, un groupe composé de deux variables très corrélées sera presque équivalent à un groupe d'une seule variable, alors qu'un groupe composé de deux variables non corrélées conservera une inertie égale à 1 dans les deux directions orthogonales, et donc une inertie totale égale à 2.

Selon les aspects de la méthode proposée que l'on considère, cette pondération sert toujours le même objectif mais de manière différente :

a) Dans la représentation simultanée, elle intervient dans la définition du nuage moyen : le rôle maximum d'un nuage $\mathcal{N}_I^j$ dans une direction donnée est égalisé. Bien entendu, si un nuage a plus de dimensions significatives que les autres, il influera sur plus de dimensions du nuage moyen. La pondération joue aussi un rôle direct sur la représentation des $\mathcal{N}_I^j$. En effet, cette pondération se traduit dans la métrique M de $\mathbf{R}^K$, et sur les nuages par une homothétie. Elle a, sur ces nuages, un effet de normalisation (basée sur la direction de plus grande inertie) qui facilite la comparaison.

b) Si l'on considère l'analyse en composantes principales de l'ensemble des variables en elle même, la pondération équilibre le rôle des différents groupes dans la détermination des facteurs. La détermination d'un facteur s'appuyant seulement sur l'inertie dans une direction donnée, il est naturel de normaliser cette dernière pour équilibrer le rôle des différents groupes.

c) Dans l'analyse des liaisons (analyse multicanonique). cette pondération rend égale à 1 la liaison maximum entre une variable et un groupe (la variable la plus liée au groupe est la première composante principale du groupe). Il est clair que, de ce point de vue encore, la pondération équilibre le rôle des différents groupes. En outre, la valeur de cette liaison devient interprétable puisqu'elle est comprise entre zéro et un.

d) Dans la comparaison globale des groupes, la contrainte que nous imposons sur le sous-espace sur lequel nous projetons les W_j est telle que, seule l'inertie des nuages dans une direction donnée joue un rôle dans la détermination du sous espace. Il est donc normal, pour équilibrer le rôle des groupes, d'égaliser l'inertie maximum dans une direction donnée. Dans l'optique du modèle INDSCAL, les facteurs communs étant normés, le poids maximum pouvant être effecté à un facteur par un groupe est égal à 1. (le facteur serait alors la première composante principale du groupe). Comme pour la mesure de liaison, la borne supérieure du poids fixée permet d'interpréter directement la valeur des poids.

8. CAS DES VARIABLES QUALITATIVES

Jusqu'ici nous n'avons parlé que d'analyse en composantes principales en nous limitant implicitement au cas de variables numériques. Or les idées présentées peuvent être appliquées aux variables qualitatives.

Une variable qualitative est équivalente à une partition de l'ensemble des individus. Elle est représentée par l'ensemble des variables indicatrices des classes de cette partition ou par le sous espace de $\mathbf{R}^I$ qu'elles engendrent. Un ensemble de variables qualitatives est ainsi codé sous forme d'un tableau disjonctif complet.

Ce type de tableau est traité classiquement par l'analyse des correspondances, appelée alors analyse des correspondances multiples. Les facteurs sur les individus pourraient être obtenus par une analyse en composantes principales, à condition de pondérer chaque variable indicatrice de manière à égaliser leur inertie. Cette pondération est indispensable pour respecter la structure d'une variable qualitative.

Il faut considérer deux cas distincts dans l'application de notre méthode aux variables qualitatives.

a) Dans le premier, un groupe est formé par les variables indicatrices définissant une variable qualitative (àvec les pondérations évoquées ci-dessus). Appliquée à cette situation, la méthode que nous venons de présenter conduit exactement aux résultats de l'analyse des correspondances multiples : les groupes de variables ont tous le même poids ; les facteurs du nuage moyen sont les facteurs définis en A.F.C. sur l'ensemble des individus ; les représentations simultanées sont homothétiques des facteurs de l'A.F.C. sur les modalités (Il faut remarquer que deux individus ayant la même modalité pour une variable sont confondus du point de vue de cette variable : les modalités suffisent pour représenter les individus vus à travers cette variable). L'analyse des liaisons se confond avec l'analyse multicanonique. Il est très satisfaisant d'obtenir, en appliquant la méthode proposée à ce cas particulier, exactement la méthode classique.

b) Dans le cas où chaque groupe de variables est constitué de plusieurs variables qualitatives, les idées présentées s'appliquent en remplaçant l'analyse en composantes principales du groupe K_j par une analyse des correspondances.

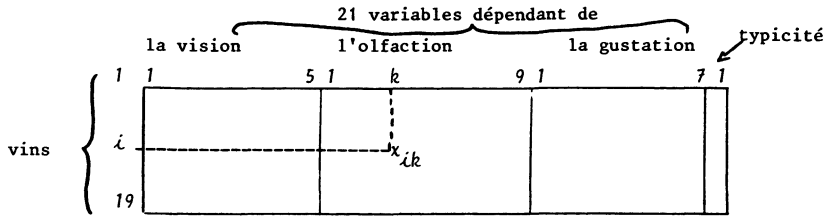
9. EXEMPLE COMMENTE

Rappelons que le tableau étudié contient les moyennes des notes attribuées par des dégustateurs à 19 vins pour 22 critères. Ces critères concernent les 3 sens en jeu dans de telles dégustations :

5 concernent la vision (intensité plus ou moins forte de la couleur, nuance plus ou moins violette. . .)
9 concernent l'ofaction (intensité plus ou moins forte du bouquet par voie directe, et par voie dite "rétronasale", aspect plus ou moins fruité du bouquet. . .)
7 concernent la gustation (acidité, astringence. . .)
1 critère est d'ordre général : la typicité. Nous l'introduirons en tant que variable supplémentaire.

Ces données sont issues d'une vaste étude dont les objectifs sont d'avoir une caractéristation rigoureuse du milieu viticole et d'étudier l'influence des différentes composantes sur la qualité et la typicité des vins. Cette étude a été réalisée par C. ASSELIN et R. MORLAT à l'INRA d'Angers [7].

En résumé, le tableau analysé possède la structure suivante :



x_{ik} : moyenne des notes attribuées par les 24 dégustateurs au vin i concernant la variable k.

Allure générale du tableau de données

9.1. La pondération des groupes

La pondération des groupes nécessite l'analyse en composantes principales des trois sous-tableaux. L'inertie de la première composante est respectivement :

3.65 pour la vision (pourcentage d'inertie 72.9) ;
5.61 pour l'olfaction (pourcentage d'inertie 62.3) ;
5.31 pour la gustation (pourcentage d'inertie 75.8).

Les poids affectés aux variables des groupes olfaction et gustation sont à peu près analogues (bien que le nombre de variables de ces deux groupes diffère). Celui qui est affecté aux variables de la vision est plus élevé.

9.2. Interprétation des facteurs

Nous interprétons d'abord l'analyse en composantes principales de l'ensemble des variables ainsi pondérées qui sert de base à tous les autres résultats.

Nous limiterons l'interprétation au premier plan factoriel qui représente 75 % de l'inertie, le premier facteur à lui seul, avec une inertie de 2.71, en représente 63 %.

Le premier facteur

La plupart des variables sont fortement corrélées positivement avec ce facteur. Seuls font exception la limpidité, l'aspect floral ou épicé du bouquet, la qualité des arômes appréciée par voie directe et l'acidité. On appelera ce premier facteur "puissance du vin", car la puissance apparaît comme le dénominateur commun des variables qui lui sont très corrélées.

Le deuxième facteur

Il oppose essentiellement le vin n° 10 (celui-ci contribue à 67 % de l'inertie de l'axe) aux autres. Les liaisons entre variables exprimées par ce second facteur sont dues presque exclusivement à ce vin.

Comparaison globale des groupes de variables

Pour comparer les groupes de variables on dispose de plusieurs indices basés sur l'analyse précédente. Il s'agit essentiellement des contributions des groupes de variables. Elles sont résumées dans le tableau suivant :

Contributions des 3 groupes aux 2 premiers facteurs

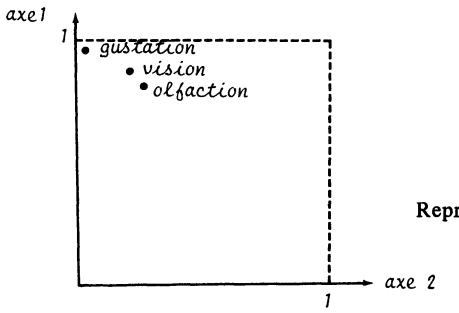
Groupe	Facteur 1	Facteur 2
vision	0.894	0.201
olfaction	0.846	0.285
gustation	0.972	0.024

Ces contributions possèdent plusieurs significations.

a) Elles montrent l'importance relative de chacun des 3 groupes dans la détermination des facteurs, (de ce point de vue elles apparaissent comme la somme des contributions des variables du groupe). Ici, les trois contributions sont à peu près les mêmes pour le premier facteur. Par contre, en ce qui concerne le deuxième facteur la contribution du groupe gustation est très inférieure aux autres.

b) Ces contributions expriment la liaison entre le groupe de variables et le facteur. Rappelons que la valeur maximum de cette liaison est 1. Ici, seul le premier facteur est très lié au trois groupes de variables. Le second facteur est légèrement lié aux groupes "vision" et "olfaction" mais pas du tout au groupe "gustation". D'ailleurs, la première composante principale de chacun des 3 groupes est très corrélée avec le premier facteur ; par contre la seconde n'est faiblement corrélée qu'avec les seconds facteurs de la vision et de l'olfaction.

c) Ces contributions (cf. paragraphe 6) sont les coordonnées des points représentant les groupes dans R^{I^2} sur les axes associés aux facteurs de l'ACP. Elles conduisent au graphique suivant :



Représentation des groupes de variables dans R^{I^2}

Ces trois points sont bien représentés sur le 1er axe et à fortiori sur le plan comme l'indique le tableau ci-dessus.

Qualité de représentation des groupes de variables dans R^{I^2}

Groupe	Facteur 1	Facteur 2
vision	0.744	0.037
olfaction	0.650	0.073
gustation	0.913	0.000

Il est donc possible de juger de la proximité de ces groupes par la proximité de leur projection sur ce plan. Ainsi, on peut dire que globalement les 3 groupes sont très proches et induisent des structures analogues sur les vins. Toutefois,

- du point de vue du premier axe le groupe "vision" est intermédiaire entre les deux autres ;
- du point de vue du deuxième axe les groupes "vision" et "olfaction" sont très proches.

d) Le graphique précédent peut être interprété en terme de modèle INDSCAL, de la façon suivante :

– les axes correspondent aux facteurs communs ;
– les coordonnées sont les poids. Appliqués à ces données, ces poids correspondent à l'influence des facteurs dans l'appréciation des vins par les trois sens.

Ici, le poids du premier facteur est important et sensiblement du même ordre pour les trois sens. Remarquons toutefois que c'est pour la gustation que ce poids est le plus grand. En outre la qualité de représentation du groupe "gustation" par ce facteur est très bonne (0,913) : ce facteur résume bien à lui seul l'ensemble des variables de ce groupe. Ce phénomène est certainement à relier avec le caractère fruste de la "gustation" comparée aux deux autres sens. Par contre cette qualité de représentation est moins bonne pour les deux autres groupes : des différences entre les vins, indépendantes de sa puissance sont ressenties par l'œil et par le nez. Une faible partie de ces différences est exprimée par le deuxième facteur, mais nous verrons ci-dessous que les représentations simultanées des vins à travers la vision et l'olfaction sont si différentes pour ce facteur, qu'il est difficile de l'interpréter comme un facteur commun.

9.3. Représentation simultanée des vins vus par les trois sens

Rappelons que dans cette représentation, chaque vin figure quatre fois : trois fois en tant que vin perçu au moyen de l'un des sens et une fois en tant que vin perçu globalement. Pour chaque vin ce dernier point figure au centre de gravité des trois premier.

Premier facteur

Les quatre points représentant le même vin sont généralement proches entre eux. Ce fait, qui peut être aisément apprécié par simple consultation visuelle du plan factoriel, peut aussi s'exprimer au travers d'un indice. En effet, en appelant :

- inertie totale, l'inertie des 3 x 19 points représentant les vins appréciés à l'aide d'un seul sens ;
- inertie inter, l'inertie de leur 19 centres de gravité (vins appréciés à l'aide des trois sens).

Le rapport (inertie inter)/(inertie totale) exprime globalement la proximité des points représentant le même vin. Pour le premier facteur, ce coefficient vaut 0,914.

Ainsi, les échelles de puissance induites sur les vins par les trois sens sont analogues. Ce sont souvent les exceptions à la règle générale qui sont intéressantes :

a) le vin 10 a paru très puissant à l'œil et moins que moyen au nez.

b) les numéros 7 et 17 représentaient le même vin témoin proposé en début et en fin de dégustation. Globalement, ce vin parait moins puissant lorsqu'il est

goûté en fin de séance, ce qui traduit un phénomène de fatigue des sens. La représentation simultanée nous permet de préciser que cette lassitude ne s'exprime qu'au niveau du goût.

La qualité de représentation des nuages $\mathcal{N}_I^j$ est mesurée par deux indices différents. Le premier, très pessimiste, se réfère à la projection effectivement réalisée sur un axe u de $\mathbf{R}^k$.

$$\text{Premier indice} = \frac{\text{inertie de } \mathcal{N}_I^j \text{ projetée sur u}}{\text{inertie totale de } \mathcal{N}_I^j}$$

Le second se réfère à sa projection sur la projection u_j de u sur $\mathbf{R}^{Kj}$ (remarque 2 du § 3).

$$\text{Deuxième indice} = \frac{\text{inertie de } \mathcal{N}_I^j \text{ projetée sur } u_j}{\text{inertie totale de } \mathcal{N}_I^j}$$

Le deuxième indice ne s'additione pas suivant les axes et présente de l'intérêt pour le premier facteur essentiellement.

Qualité de représentation des 3 nuages $\mathcal{N}_I^j$ sur le premier facteur

Groupe	Projection sur u	Projection sur u_j
vision	0.238	0.720
olfaction	0.189	0.605
gustation	0.271	0.755

La qualité de représentation des trois nuages sur les u_j est donc très proche de celle du premier facteur de l'analyse de ces groupes. Ceci n'est pas étonnant puisque les premiers facteurs de ces nuages sont très liés au premier facteur du nuage moyen.

Deuxième facteur

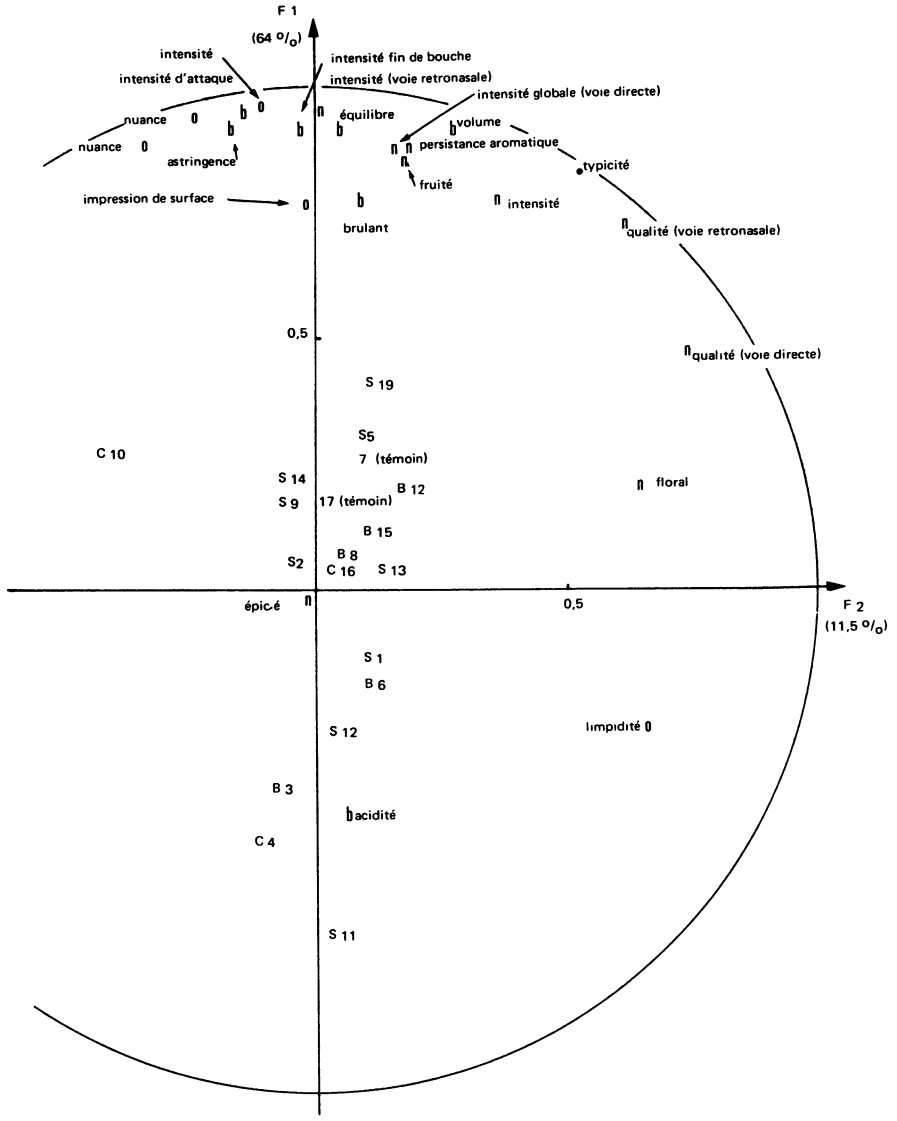
La situation est très différente de celle du premier facteur. Les points représentant le même vin sont généralement éloignés. Cette constatation visuelle est étayée par la valeur de l'indice (Inertie inter-inertie totale) qui ne vaut que 0.288. Nous ne nous attacherons donc pas à l'interprétation de cet axe.

10. CONCLUSION

Cet exemple réel montre que la méthode proposée permet de répondre aux questions posées par le traitement de tableaux croisant un ensemble d'individus et plusieurs groupes de variables. Les différents éclairages donnés aux résultats les précisent et les enrichissent considérablement.

Le volume de ces résultats est important mais ne nécessite pas pour autant de lourds calculs : à peine plus qu'une simple analyse en composantes principales du tableau qui en constitue, d'ailleurs, la part essentielle. Le simple fait que l'exemple ait pu être traité sur Mini 6 le prouve.

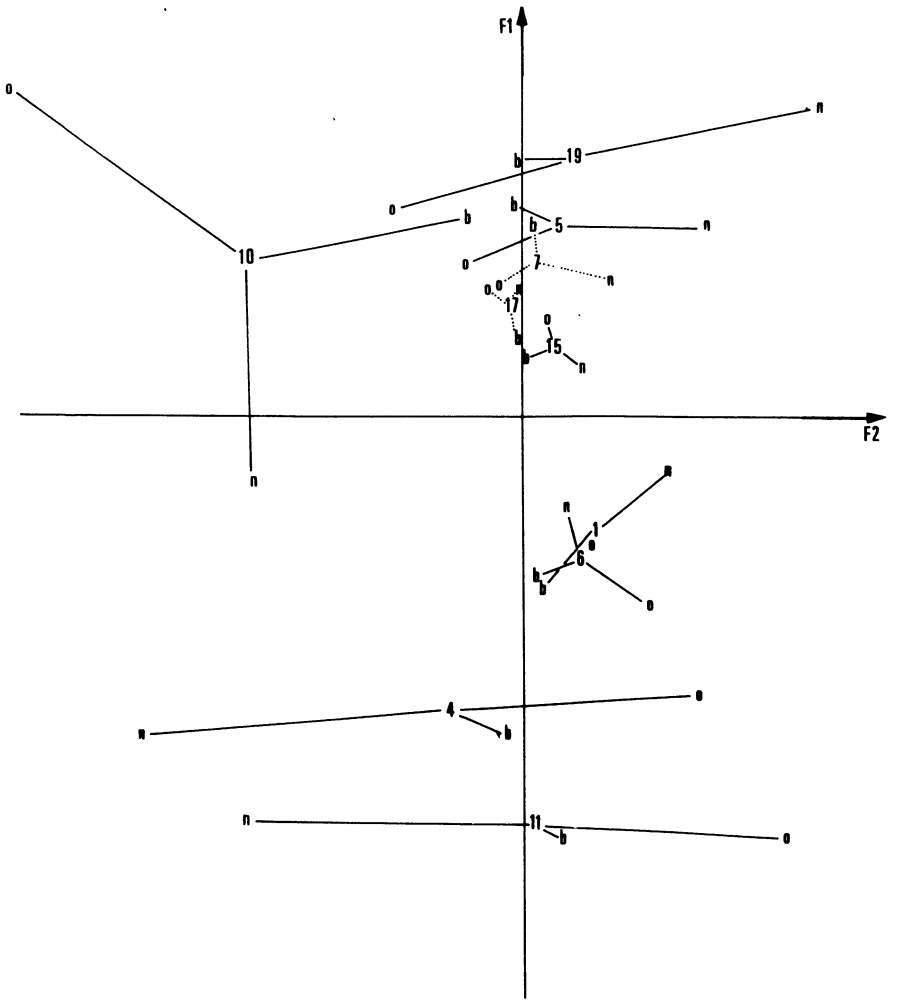
Cette méthode ne s'applique pas directement à l'étude d'autres types de tableaux structurés, croisant par exemple un ensemble de variables et plusieurs groupes d'individus. (les notions de ressemblance entre groupes de variables et entre groupes d'individus sont de natures différentes). Nous proposerons ultérieurement les variantes de cette méthode adaptées à d'autres types de tableaux structurés.



Graphique 1. – Caractérisation de vins rouges du Val-de-Loire ; représentation des variables et des individus moyens.

Symboles utilisés :

Pour les variables	Pour les individus
o = œil	S = Saumur
n = nez	C = Chinon
b = bouche	B = Bourgueil



Graphique 2. – Caractérisation de vins rouges du Val de Loire : représentation simultanée des vins vus par chacun des groupes de variables.

Le numéro du vin figure au point moyen. Il est relié à ses points homologues o (œil), n (nez), et b (bouche). Par souci de clarté, seuls les vins extrêmes ou donnant lieu à un commentaire dans le texte ont été représentés.

BIBLIOGRAPHIE

[1] J.D. CARROLL. – *A generalization of canonical correlation analysis to three or more sets of variables.* Proceedings of the 76th annual convention of the American Psychological association, p. 227-228, 1968.

[2] J.D. CARROLL et J.J. CHANG. – Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart Young" decomposition. *Psychometrika,* vol. 35 n° 3, p. 283-319, 1970.

[3] B. ESCOFIER et J. PAGES. – Comparaison de groupes de variables définies sur le même ensemble d'individus. *Rapport IRISA,* Mai 1982 n° 166. Comparaison de groupes de variables (2): Application à la caractérisation de vins rouges du Val de Loire. *Rapport IRISA,* n° 172, Juillet 82.

[4] Y. ESCOUFIER. – L'analyse conjointe de plusieurs matrices. *Biométrie et temps. Société Française de Biométrie,* 1980.

[5] J.R. KETTENRING. – Canonical analysis of several sets of variables. *Biometrika,* Vol. 58 n° 3, p. 433-451, 1976.

[6] H. L'HERMIER DES PLANTES (1976). – *Structuration des tableaux à trois indices de la statistique.* Thèse de 3e cycle, Université de Montpellier, 1976.

[7] R. MORLAT et C. ASSELIN. – *Etude de l'Influence du milieu sur la qualité et la typicité du vin. L'essai terroir.* Publication interne, Centre de Recherche INRA Angers, 1981.