REVUE DE STATISTIQUE APPLIQUÉE

F. HUSSON
J. PAGÈS

Nuage plan d'individus et variables supplémentaires

Revue de statistique appliquée, tome 51, n° 4 (2003), p. 83-93
<http://www.numdam.org/item?id=RSA_2003__51_4_83_0>

Rev. Statistique Appliquée, 2003, **LI** (4), 83-93

NUAGE PLAN D'INDIVIDUS ET VARIABLES SUPPLÉMENTAIRES

F. HUSSON, J. PAGÈS

Laboratoire de Mathématiques appliquées, ENSA de Rennes,
65 rue de Saint-Brieuc, 35042 Rennes Cedex
husson@agrorennes.educagri.fr

RÉSUMÉ

On veut souvent compléter l'interprétation des graphes usuels (x, y) à l'aide de variables quantitatives supplémentaires. La méthode PREFMAP (modèle vectoriel) propose une représentation de ces variables supplémentaires. Dans cet article, on met en évidence quelques inconvénients de la méthode PREFMAP lorsque les variables x et y sont corrélées et l'on propose une représentation analogue en substituant dans la méthode PREFMAP la régression usuelle par la régression PLS. Le graphique obtenu est enrichi par une visualisation de la qualité de représentation des variables à l'aide de courbes de niveau.

Mots-clés : PREFMAP, PLS, biplot

ABSTRACT

Often one wants to complete the interpretation of the usual graphs (x, y) with additional quantitative variables. The PREFMAP method (vectorial model) proposes a representation of these additional variables. In this article, we highlight some disadvantages of the PREFMAP method when the variables x and y are correlated and we propose a similar representation to the PREFMAP method but with the substitution of the regression by the PLS regression. The graph obtained is enriched by a representation of the level lines of quality of representation.

Keywords : PREFMAP, PLS, biplot

1. Introduction

La représentation d'un nuage de n individus en fonction de deux variables quantitatives (x et y) est sans doute le graphe le plus utilisé en statistique. L'interprétation de tels graphes est enrichie lorsque l'on peut relier x et y à d'autres variables disponibles. Notons $\mathbb{R}^n$ l'espace des variables et notons $E_{x,y}$ le plan de $\mathbb{R}^n$ engendré par les variables x et y. Le paradigme est le graphe de l'ACP où les variables x et y sont les composantes principales. Soit k une variable centrée réduite (nous supposerons dans la suite de cet article que les variables sont centrées réduites, ce qui ne nuit pas à la généralité des résultats obtenus). Le graphe de l'ACP permet de juger la corrélation de la variable k avec chacune des composantes principales (la coordonnée de la variable

k sur l'axe s est son coefficient de corrélation avec la composante principale s) mais également de juger la liaison entre la variable k et l'ensemble des 2 composantes principales (grâce à la proximité entre le point et le cercle des corrélations). Cette propriété de l'ACP est assurée par la non-corrélation des composantes principales. Ce graphe présente donc des propriétés intéressantes :

- les coordonnées d'une variable k correspondent au coefficient de corrélation entre k et chacune des composantes principales (si k est centrée réduite);
- la norme (au carré) du projeté de k sur le plan correspond au coefficient de détermination de la régression de k en fonction des variables x et y;
- le coefficient de corrélation entre deux variables k et l correspond au cosinus de l'angle entre k et l dans l'espace $\mathbb{R}^n$. Si les variables k et l sont proches du plan de projection $E_{x,y}$, alors l'angle entre le projeté de k et le projeté de l est proche de l'angle entre k et l et donc le cosinus de l'angle entre le projeté de k et le projeté de l donne une bonne idée du coefficient de corrélation entre k et l. Si les variables sont mal représentées, on ne peut rien dire concernant leur liaison;
- il y a une dualité entre la représentation des individus et celle des variables et donc possibilité de faire une lecture simultanée du graphe des individus et du graphe des variables.

Cependant, les variables que l'on étudie x et y sont très souvent corrélées. C'est presque toujours le cas lorsqu'elles sont définies *a priori*, mais c'est également le cas lorsqu'elles sont issues d'analyse comme, par exemple, la construction d'un modèle INDSCAL (Carroll et Chang, 1970). Ce problème a déjà été étudié et est classiquement résolu par l'aspect vectoriel de la méthode PREFMAP (Caroll, 1972, 1980). On présente alors un graphe avec x et y orthogonaux (on notera ce plan $\mathbb{R}^2_{x,y}$) et on représente chaque variable k de la façon suivante : ses coordonnées sont proportionnelles aux coefficients de la régression de k en fonction des régresseurs x et y et sa norme est égale à la racine carrée du coefficient de détermination de cette régression. La norme mesure ici la qualité de représentation de la variable sur le plan $\mathbb{R}^2_{x,y}$: si la norme du projeté de la variable est proche de 1, alors le coefficient de détermination est proche de 1 donc la variable est proche du plan de projection.

Cette méthode fondée sur la régression en présente les inconvénients : lorsque les prédicteurs sont corrélés, les coefficients de la régression ne représentent pas la liaison «directe» entre la variable et le prédicteur. Il est même possible qu'un coefficient soit négatif alors que le coefficient de corrélation entre le prédicteur et la variable est positif.

On propose dans cet article une représentation analogue à celle de PREFMAP mais en remplaçant la régression linéaire par la régression PLS-1 ce qui résout les problèmes d'interprétation des coefficients de la régression.

2. Méthode pour une représentation de variables associée à un nuage plan d'individus

2.1. Principe

Dans la forme la plus simple de la régression PLS-1 (nous utilisons ici la terminologie de Tenenhaus où une seule variable est à expliquer) on utilise une seule composante PLS : les coefficients des prédicteurs dans l'équation de prédiction sont, à un facteur multiplicatif près, les coefficients de corrélation entre la variable et les prédicteurs. Remplacer dans PREFMAP la régression usuelle par la régression PLS-1 revient alors à construire un graphe similaire à celui de l'ACP puisque chaque coordonnée correspond au coefficient de corrélation entre le facteur et le prédicteur. Ces graphes sont utilisés par exemple dans la procédure *bivar* de Spad (SPAD, 2001). Dans ce graphique, les coordonnées des variables sont interprétables mais leur distance à l'origine ne l'est pas puisque deux variables de même longueur peuvent avoir des qualités de représentation différentes. La relation entre la longueur d'une variable et sa qualité de représentation dépend de la corrélation entre les variables x et y et, à corrélation entre x et y fixée, de la direction de la variable. D'où l'idée de construire des courbes de niveau de qualité de représentation dont la plus extrême serait l'analogue du cercle des corrélations en ACP.

2.2. Domaine de définition de la variable k dans le plan $\mathbb{R}^2_{x,y}$

La qualité de la représentation de la variable k dans le plan $\mathbb{R}^2_{x,y}$ (proposée dans le paragraphe 2.1) peut être mesurée par le coefficient de corrélation entre la variable k et sa représentation sur $\mathbb{R}^2_{x,y}$. Ce coefficient de corrélation (que l'on notera R) est fonction des coefficients de corrélation entre k et x (*i.e.* r_{kx}) et entre k et y (*i.e.* r_{ky}) puisque la coordonnée de k sur l'axe x (resp. y) correspond à r_{kx} (resp. r_{ky}).

Déterminons dans un premier temps le domaine de définition de la fonction f qui au couple (r_{kx}, r_{ky}) associe R. f est une fonction de $[-1;1]^2$ dans $[0;1]$. Cependant, les coefficients de corrélation r_{kx} et r_{ky} ne sont pas indépendants. Ils dépendent du coefficient de corrélation r_{xy}.

Le coefficient de corrélation r_{xy} étant fixé, on peut calculer les bornes du domaine de définition de f en considérant fixée la valeur de r_{kx}. Il suffit alors de borner les valeurs que peut prendre r_{ky}, ce qui est réalisé par la proposition 2.1.

PROPOSITION 2.1. – *Le domaine de définition de la fonction f est tel que* :

$$r_{xy}r_{kx} - \sqrt{(1-r_{xy}^2)(1-r_{kx}^2)} \leqslant r_{ky} \leqslant r_{xy}r_{kx} + \sqrt{(1-r_{xy}^2)(1-r_{kx}^2)} \quad (1)$$

Preuve. – Les trois variables x, y et k étant dans un espace à au plus trois dimensions, la matrice de corrélation des trois variables x, y, et k doit être définie positive ou semi-définie positive. Or ceci est vrai si et seulement si le déterminant de cette matrice est positif ou nul (l'implication est vraie par définition, la réciproque est vraie car toutes les corrélations sont inférieures à 1). On a alors :

$$\det\begin{pmatrix} 1 & r_{xy} & r_{kx} \\ r_{xy} & 1 & r_{ky} \\ r_{kx} & r_{ky} & 1 \end{pmatrix} = 1 + 2\, r_{xy} r_{kx} r_{ky} - (r_{xy}^2 + r_{kx}^2 + r_{ky}^2)$$

Et ce déterminant est positif ou nul si et seulement si

$$r_{xy}r_{kx} - \sqrt{(1-r_{xy}^2)(1-r_{kx}^2)} \leqslant r_{ky} \leqslant r_{xy}r_{kx} + \sqrt{(1-r_{xy}^2)(1-r_{kx}^2)}$$

La figure 1 donne les limites du domaine de définition (le domaine de définition correspond à l'intérieur des ellipses) pour des coefficients de corrélation r_{xy} variant entre 0 et 0.8. On peut remarquer que pour un coefficient de corrélation de 0, on retrouve bien le cercle des corrélations de l'ACP. Lorsque le coefficient de corrélation r_{xy} est élevé, la forme du domaine de définition devient une ellipse d'autant plus « allongée» que r_{xy} est grand.

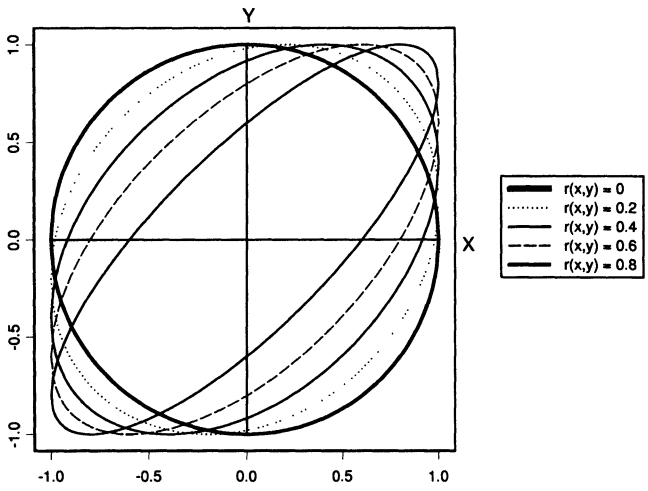

FIGURE 1
Limite du domaine de définition de R pour plusieurs coefficients de corrélation entre x et y

2.3. Qualité de représentation

PROPOSITION 2.2. – *Sur le domaine de définition où la fonction f est définie (voir proposition 2.1), on a* :

$$R = \frac{r_{kx}^2 + r_{ky}^2}{\sqrt{r_{kx}^2 + r_{ky}^2 + 2 \times r_{kx} \times r_{ky} \times r_{xy}}} \qquad (2)$$

Preuve. – Par construction, la représentation de la variable k sur $\mathbb{R}^2_{x,y}$ correspond à la prédiction de k par la méthode PLS-1 (à une normalisation près par le coefficient multiplicateur précité). On notera alors ce projeté $\hat{k}_{pls}$ et on adoptera la norme Id_n/n dans $\mathbb{R}^n$ (Id_n étant la matrice Identité de taille $n \times n$). Ainsi on a

$$\hat{k}_{pls} = XX'k/n$$

avec X la matrice ayant pour première colonne les valeurs prises par la variable x et ayant pour deuxième colonne les valeurs prises par y.

R étant le coefficient de corrélation entre k et sa représentation sur $\mathbb{R}^2_{x,y}$ (*i.e.* $\hat{k}_{pls}$), on a :

$$R(k, \hat{k}_{pls}) = \frac{k'\hat{k}_{pls}/n}{\|k\| \, \|\hat{k}_{pls}\|} = \frac{k'XX'k/n^2}{1 \times \|XX'k/n\|}$$

$$R(k, \hat{k}_{pls}) = \frac{r^2_{kx} + r^2_{ky}}{\|XX'k/n\|}$$

La norme de $XX'k/n$ correspond à la longueur de la diagonale d'un parallélogramme ayant des côtés de longueur r_{kx} et r_{ky} et formant un angle $\widehat{(x,y)}$, et donc :

$$\begin{aligned} \|XX'k/n\|^2 &= (r_{kx} + r_{ky} \times \cos(x,y))^2 + (r_{ky} \times \sin(x,y))^2 \\ &= r^2_{kx} + r^2_{ky} + 2 \times r_{kx} \times r_{ky} \times \cos(x,y) \\ &= r^2_{kx} + r^2_{ky} + 2 \times r_{kx} \times r_{ky} \times r_{xy} \quad \text{car } \cos(x,y) = r_{xy} \end{aligned}$$

Ainsi, on a

$$R = \frac{r^2_{kx} + r^2_{ky}}{\sqrt{r^2_{kx} + r^2_{ky} + 2 \times r_{kx} \times r_{ky} \times r_{xy}}} \tag{3}$$

2.4. Courbes de niveau de qualité de représentation

Les graphes de la figure 2 montrent l'allure des courbes de niveau de qualité de représentation pour des corrélations entre x et y variant de 0 à 0.9; les graphes obtenus avec des corrélations négatives ne sont pas présentés car ce sont les symétriques, par rapport à l'axe des ordonnées, de ceux obtenus lorsque la corrélation est positive. Quand les deux facteurs sont indépendants, on retrouve le cercle de corrélation de l'ACP. Lorsque r_{xy} est grand en valeur absolue, les courbes de niveau de qualité de représentation sont indispensables pour connaître la qualité de représentation d'une

variable. En effet, il est clair que la distance d'un point à l'origine du graphe ne représente pas la qualité de représentation.

3. Application : projection de variables supplémentaires dans le cadre d'un modèle INDSCAL

Le modèle INDSCAL (Carroll et Chang, 1970) permet d'effectuer une évaluation directe et globale de la différence sensorielle entre produits. Le protocole classique consiste à proposer à un ensemble d'individus (appelés juges) un ensemble de produits, paire par paire. Chaque juge évalue la différence entre deux produits par une note et fournit finalement une matrice d'indices de dissimilarités.

Ainsi, pour J juges et I produits, le juge j évalue globalement la dissimilarité entre les produits i et l par la note $d_j(i,l)$ variant par exemple de 0 (produits identiques) à 10 (produits très différents). Selon le modèle INDSCAL, la dissimilarité $d_j(i,l)$ dérive d'une configuration des produits en S dimensions (soit $z_s(i)$ la coordonnée du produit i sur l'axe de rang s de cette configuration), chaque juge j accordant un poids spécifique q_s^j à la dimension s :

$$d_j^2(i,l) = \sum_{s=1}^{S} q_s^j \left(z_s(i) - z_s(l)\right)^2 + e_j(i,l) \tag{4}$$

où $e_j(i,l)$ est le résidu du modèle.

L'estimation des paramètres de ce modèle est effectuée de manière itérative (une itération comprenant une estimation des coordonnées $z_r(i)$ et une estimation des poids). Les facteurs ainsi obtenus sont souvent corrélés.

Nous appliquons la procédure présentée ci-dessus à des sorties du modèle INDSCAL dérivant d'une application en analyse sensorielle. Dans cette application, six chocolats ont été évalués par six jurys à l'aide de 14 descripteurs (*i.e.* variables) sensoriels. À partir des notes obtenues sur les 14 descripteurs, les dissimilarités entre chocolats ont été calculées (par la distance euclidienne). Une description détaillée du plan d'expériences et des données est fournie dans l'article de Pagès et Husson (2001). L'analyse de ces données dans le cadre du modèle INDSCAL fournit une configuration «commune» des six chocolats (voir figure 3) dans laquelle les deux facteurs F_1 et F_2 sont fortement corrélés ($r_{F_1,F_2} = -0.805$). La construction du modèle INDSCAL est détaillée et commentée dans l'article de Pagès et Husson (soumis à *Computational Statistics and Data Analysis*). Pour interpréter cette configuration «commune», on dispose de trois variables physico-chimiques (la teneur en CACAO, le pourcentage de *SACCHAROSE* et le pourcentage de *MATIÈRE GRASSE*) ainsi que des moyennes (tout jury confondu) des 14 descripteurs sensoriels (*acide*, *amer*, *sucré*, etc.).

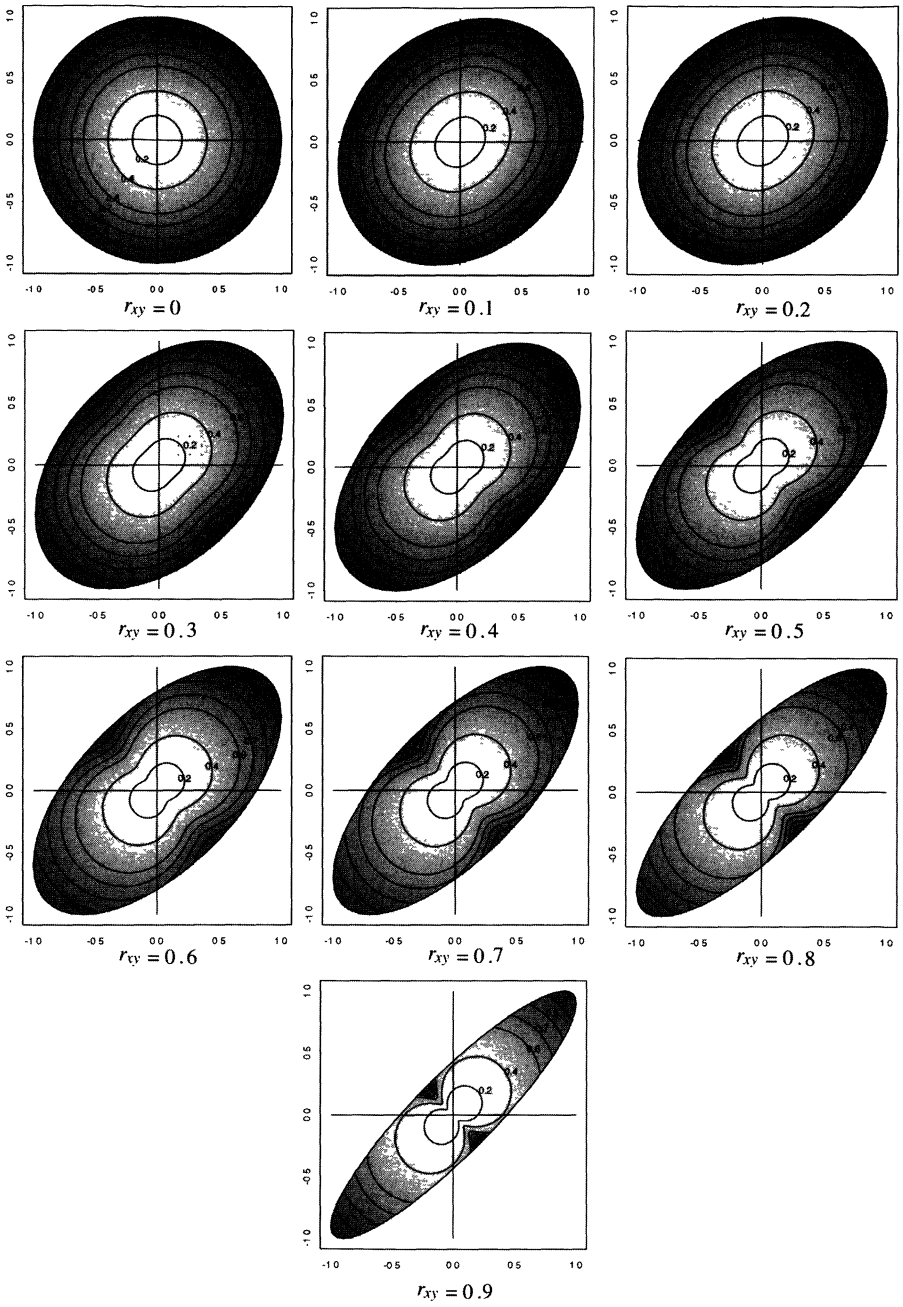

FIGURE 2
Courbes de niveau de qualité de représentation pour plusieurs coefficients de corrélation entre x et y

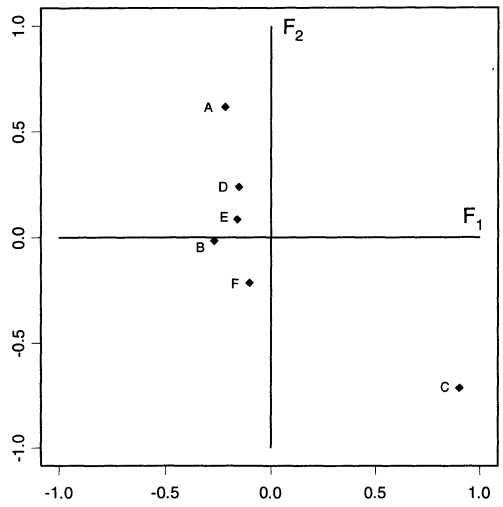

FIGURE 3
Représentation des chocolats obtenue par le modèle INDSCAL
($r_{F_1,F_2} = -0.805$)

3.1. Corrélation entre les variables

Méthode PREFMAP **(figure 4).** Le graphe suggère une liaison entre les variables *fondant*, *O.lait* et *S.lait* d'une part et le facteur 1 d'autre part. Effectivement, ces trois variables sont étroitement corrélées avec ce facteur ($r_{fondant,F_1} = 0.995$, $r_{O.lait,F_1} = 0.995$ et $r_{S.lait,F_1} = 0.994$). En revanche, ce graphe suggère une absence de liaison entre les variables (teneur en) *CACAO* et *acide* d'une part et le facteur 1 d'autre part. Ceci n'est pas vrai puisque $r_{CACAO,F_1} = -0.754$ et $r_{acide,F_1} = -0.764$. *A fortiori*, le graphe suggère une absence de liaison entre les variables *acide* et *fondant* (par exemple) alors que $r_{acide,fondant} = -0.700$.

Méthode PREFMAP-PLS **(figure 5).** Par définition, on lit bien sur le graphe les corrélations entre les variables et chacun des facteurs. La plupart des variables sont fortement corrélées aux deux facteurs. Les courbes de niveau indiquent les corrélations qui peuvent être déduites de la position des variables. Ainsi, par exemple, *sucré* et *vanille* sont bien représentées et leur proximité suggère une forte corrélation. En effet, $r_{sucré,vanille} = 0.98$. En revanche, *SACCHAROSE* et *croquant* sont mal représentées et on ne peut rien inférer de leur position relative. Leur opposition sur le graphe ne traduit pas forcément une corrélation négative. En effet, $r_{SACCHAROSE,croquant} = 0.06$.

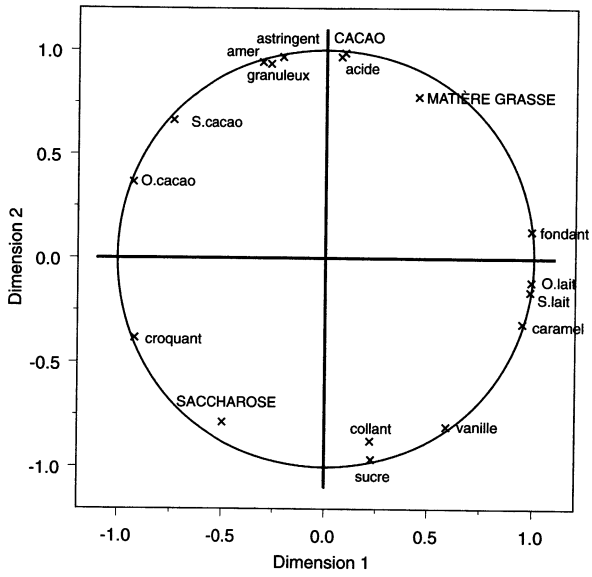

FIGURE 4
Représentation des variables par la méthode PREFMAP
sur les données chocolat ($r_{F_1,F_2} = -0.805$)

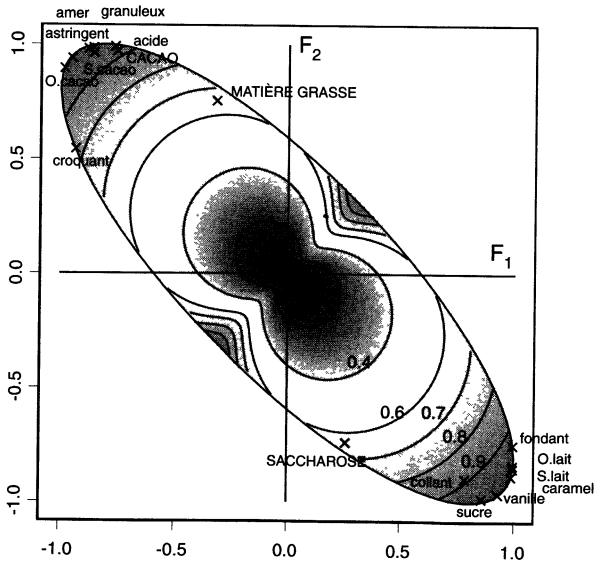

FIGURE 5
Représentation des variables sur les données chocolat ($r_{F_1,F_2} = -0.805$)

TABLEAU 1

Tableau des descripteurs centrés réduits

Chocolat	Odeur		Saveur											*CACAO*	*MATIÈRE*	*SACCHAROSE*
	cacao	*lait*	*sucré*	*acide*	*amer*	*cacao*	*lait*	*caramel*	*vanille*	*astringent*	*croquant*	*fondant*	*collant*	*granuleux*	*GRASSE*	
A	0.85	-0.56	-1.47	1.58	1.41	1.12	-0.65	-0.78	-1.17	1.46	-0.04	-0.36	-1.02	1.13	1.46	1.57	-1.78
B	0.55	-0.52	-0.30	-0.31	0.32	0.48	-0.47	-0.47	-0.29	0.30	1.02	-0.80	0.56	0.11	0.24	-1.03	0.08
C	-2.16	2.22	1.84	-1.68	-1.89	-2.07	2.22	2.19	2.04	-1.84	-2.08	2.20	1.80	-1.88	-1.58	-0.70	0.42
D	0.47	-0.55	-0.32	0.64	0.41	0.46	-0.49	-0.48	-0.44	0.42	0.05	-0.32	-0.55	1.01	0.49	0.27	-0.25
E	0.28	-0.42	-0.21	0.17	0.13	0.11	-0.41	-0.39	-0.37	0.12	0.33	-0.43	-1.02	0.01	0.36	0.92	-0.08
F	0.01	-0.16	0.45	-0.38	-0.37	-0.09	-0.20	-0.06	0.23	-0.46	0.72	-0.29	0.23	-0.38	-0.97	-1.03	1.61
Différence ntre A et C	-3.01	2.78	3.31	-3.26	-3.30	-3.19	2.86	2.97	3.21	-3.30	-2.04	2.57	2.82	-3.01	-3.04	-2.27	2.20

3.2. Interprétation simultanée du graphe des individus et du graphe des variables

Méthode PREFMAP-PLS **(figure 5).** Comme en ACP, dans l'interprétation simultanée du graphe des individus et des variables, on prend en compte de façon privilégiée les variables bien représentées. Ainsi, l'opposition majeure révélée par le graphique de la figure 3 (entre les chocolats A et C) concerne, certes, l'ensemble des variables mais plus particulièrement les variables bien représentées comme par exemple *sucré*, *vanille*, *astringent*, *S.cacao* et à un degré moindre les variables les moins bien représentées : *SACCHAROSE*, *MATIÈRE GRASSE* et *croquant*. Ceci peut être contrôlé par la différence entre les valeurs centrées réduites des chocolats A et C. Les différences sont toutes importantes en valeur absolue mais à un degré moindre pour les trois variables *SACCHAROSE*, *MATIÈRE GRASSE* et *croquant* (voir la dernière ligne du tableau 1).

Méthode PREFMAP **(figure 4).** Le graphique suggère que l'opposition entre d'une part *croquant* et *SACCHAROSE* et d'autre part *MATIÈRE GRASSE* ne concerne pas l'opposition entre les produits A et C. Or, le tableau 1 montre que l'opposition entre les produits A et C est bien réelle pour ces trois descripteurs même si cette opposition est plus forte pour d'autres descripteurs.

4. Conclusion

L'interprétation du graphique issu de la méthode PREFMAP n'est pas sans risque du fait de sa ressemblance avec le graphe du «cercle des corrélations». Le graphe PREFMAP-PLS annule ce risque, ne serait-ce que par sa spécificité graphique mais surtout grâce à l'information complémentaire exprimée par les courbes de niveau de qualité de représentation.

Bibliographie

Carroll J.D. (1972), *Individual Differences and Multidimensional Scaling,* in R.N. Shepard, A.K. Romney, and S.B. Nerlove (eds.), *Multidimensional Scaling : Theory and Applications in the Behavioral Sciences (Volume 1)*, New York : Seminar Press.

Carroll J.D. (1980), *Models and methods for multidimensional analysis of preferential choice (or other dominance) data* in E.D.Lantermann and H.Feger (Eds.), *Similarity and Choice*, pp. 234-289. Bern : Hans Huber.

Carroll J.D. et Chang J.J. (1970), Analysis of individual differences in multidimensional scaling via an N-way generalization of «Eckart-Young» decomposition. *Psychometrika, 35,* 283–319.

Pagès J. et Husson F. (2001), Inter-laboratory comparison of sensory profiles. Methodology and results. *Food Quality and Preference, 12,* 297–309.

Pagès J. et Husson F. (submitted) INDSCAL Model : geometrical interpretation and methodology. *Computational Statistics and Data Analysis.*

Spad (2001), *Logiciel distribué par le Decisia.*

Tenenhaus M. (1998), *La régression PLS, théorie et pratique*. Technip. Paris.