Classifications hiérarchiques factorielles de variables

Sergio Camiz**, Jean-Jacques Denimal*

** Dipartimento di Matematica Guido Castelnuovo Università di Roma La Sapienza
Piazzale Aldo Moro, 2 – I 00186 Roma Italie
sergio.camiz@uniroma1.it
http://www.camiz.net
* U.F.R. de Mathématiques Université des Sciences et Technologies de Lille
F 59655 Villeneuve d'Ascq France
jean-jacques.denimal@univ-lille1.fr

Résumé. On présente deux méthodes de classification hiérarchique ascendante de variables quantitatives et de fréquences. Chaque noeud de ces hiérarchies regroupe deux classes de variables à partir d'une analyse factorielle particulière basée sur les variables représentatives de ces deux classes. Par cette méthode, on dispose, à chaque pas, d'un plan factoriel permettant de représenter à la fois les variables des deux classes fusionnées et l'ensemble des individus. Ces derniers se positionnent dans ce plan suivant leurs valeurs pour les variables considérées. Ainsi, l'interprétation des nœuds obtenus s'effectue facilement à partir de l'examen de ces représentations factorielles. La répartition des individus observée dans chacun de ces plans factoriels permet également de définir une segmentation des individus en total accord avec la hiérarchie des variables obtenues. On montre le fonctionnement des méthodes sur des exemples réels.

1 Introduction

L'analyse exploratoire d'un tableau de données, que ce soit un tableau classique croisant unités statistiques et caractères quantitatifs, ou un tableau de contingence croisant les modalités de deux caractères qualitatifs, est généralement réalisée par les quatre étapes de la procédure suivante :

1. Analyse factorielle exploratoire : selon le type de tableau, il s'agit d'une Analyse en Composantes Principales (*ACP*) ou une Analyse des Correspondances (*AFC*) ;
2. classification des lignes, à savoir des individus ou des modalités en ligne ;
3. interprétation des classes obtenues à l'aide du comportement des caractères originaux dans chaque classe ;
4. Étude des liaisons entre classes et axes factoriels.

L'originalité de l'approche proposée dans cet article est d'unifier, dans une même méthode, l'analyse factorielle du tableau et les classifications des lignes et des colonnes. En effet, les plans factoriels obtenus sont directement associés aux nœuds des hiérarchies construites. Ce qui permet d'obtenir une interprétation conjointe des nœuds et des axes factoriels facilitant la synthèse des résultats. Les approches classiques résumées par les quatre étapes

décrites ci-dessus ne permettent pas, dans bien des cas, d'obtenir rapidement cette vue synthétique de l'ensemble des résultats En effet, les relations existant entre nœuds et axes factoriels sont souvent difficiles à expliquer dans l'approche classique, surtout pour un utilisateur peu exercé.

Le problème de la classification des variables n'a pas été beaucoup été traité en littérature, le seul exemple couramment utilisé par les non-spécialistes étant la procédure VARCLUS intégrée dans SAS (1999). Pourtant, Lerman (1981) a proposé un véritable système de méthodes de classification des caractères, intégré dans une méthodologie générale de classification.

Les méthodes qu'on propose ici, développées par Denimal (1997, 2001), se basent, comme d'ailleurs celles proposées par Qannari et al. (1999), sur des *ACP* ou, dans le cas original d'un tableau de contingence, sur des *AFC.* Nous nous distinguons, cependant, par le fait que l'on propose des méthodes hiérarchiques ascendantes, ce qui permet de retarder le choix du nombre de classes de la partition à retenir après la construction de la hiérarchie et l'interprétation des résultats. Ainsi, le nombre de classes à retenir est déterminé sur la base des résultats et non pas à priori. En outre, notre approche est en accord avec celle de Lerman pour laquelle chaque classe est également identifiée avec un facteur.

2 Les méthodes

Dans les deux cas on suppose les tableaux déjà normalisés, selon deux principes différents suivant le type de tableau. Au départ de l'algorithme, chaque colonne est vue comme un groupe singleton dont elle est la variable représentative. Ensuite la procédure itérative pour construire la hiérarchie est la suivante : à chaque pas

1. On effectue une analyse factorielle exploratoire sur tout couple de colonnes ;
2. On choisit comme noeud de la hiérarchie à construire le couple de classes dont la deuxième valeur-propre de l'analyse factorielle correspondante est minimum ;
3. Ce noeud établi, on choisit comme variable représentative du groupe ainsi formé le premier vecteur-propre de l'analyse factorielle correspondante.

On vient de remarquer que la différence entre les deux analyses consiste en la différente analyse factorielle utilisée : pour le tableau quantitatif on utilise une *ACP* non normée, à savoir basée sur la matrice de variance-covariance des deux caractères ; pour le tableau de contingence on utilise une *AFC*, mais dans ce cas on n'obtiendrait qu'un seul facteur. Donc il est nécessaire d'introduire une astuce, consistant à associer à toute colonne j une colonne j^* complément de j par rapport à la colonne marginale, à savoir telle que, pour tout i, $n_{ij^*} = n_{i.} - n_{ij}$. Cependant, pour tout couple de colonnes j, k, on effectue *l'AFC* du tableau de contingence croisant toutes les lignes avec les quatre colonnes $\{j, k, j^*, k^*\}$. Comme $j + k = j^* + k^*$, l'*AFC* ne fournit que deux facteurs non nuls. Une fois effectuée l'*AFC,* on ajoute les nouvelles colonnes j_{n1} et j_{n1}^* telles que

$$\forall i \in I \begin{cases} k(i, j_{n1}) = f_i k(., j_{n1}) \left(1 + \sqrt{\dfrac{2k(., j^*_{n1})}{k(j_{n1})}} F_1(i) \right) \\ k(i, j^*_{n1}) = f_i k(., j^*_{n1}) \left(1 - \sqrt{\dfrac{2k(., j_{n1})}{k(., j^*_{n1})}} F_1(i) \right) \end{cases}$$

où $F_1(i)$ est la coordonnée de i sur le premier facteur, f_i est la fréquence la ligne i, et k (.,j_{n1}) et k (.,j^*_{n1}) sont quantités définies par des égalités supplémentaires dépendant du signe des coordonnées de *j, k, j*,* et *k** sur le premier facteur (Denimal., 2000).

On a les propriétés suivantes :

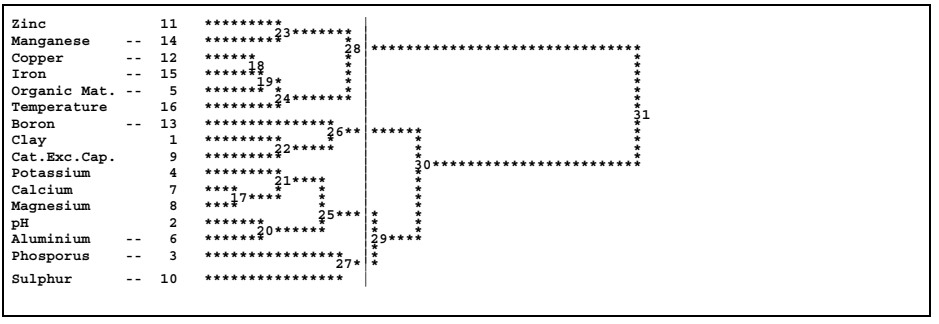
- Le critère d'agrégation est de type Ward, les deuxièmes valeurs-propres, utilisés comme indices de la hiérarchie forment une séquence non décroissante ;
- l'inertie totale du tableau est la somme des deuxièmes valeurs-propres des noeuds et de la première du nœud le plus haut ;
- à tout nœud un plan factoriel est associé, où sont représentées les colonnes des groupes fusionnés ainsi que l'ensemble des lignes ;
- dans ces plans, le premier vecteur-propre peut s'interpréter comme un compromis entre les 2 classes regroupées et le deuxième vecteur propre comme traduisant leurs différences ;
- chaque groupe s'interprète comme un dipôle, mettant en opposition des caractères ou des modalités : ceci permet une interprétation des groupes plus imagée.

3 Les exemples d'application

3.1 Données quantitatives

Dans cet exemple on a appliqué la classification hiérarchique factorielle sur un tableau de données saisies dans une communauté végétale typique des pâturages de Campos dans le Brésil du Sud (Pillar *et al.*, 1992).

Il s'agit de 60 relevés carrés de 0.5 x 0.5 m. alignés le long de quatre gradients, allant du haut en bas, dont on a la composition des 60 espèces présentes et 21 variables décrivant la structure du sol. On s'occupe ici des variables quantitatives décrivant la structure du sol. Dans le Tableau 1 on voit le dendrogramme de la hiérarchie formée par les variables et dans la Figure 1 le plan factoriel associé à l'avant-dernier nœud de la hiérarchie.



TAB. 1 – *Dendrogramme de la hiérarchie factorielle sur variables de sol de Campos.*

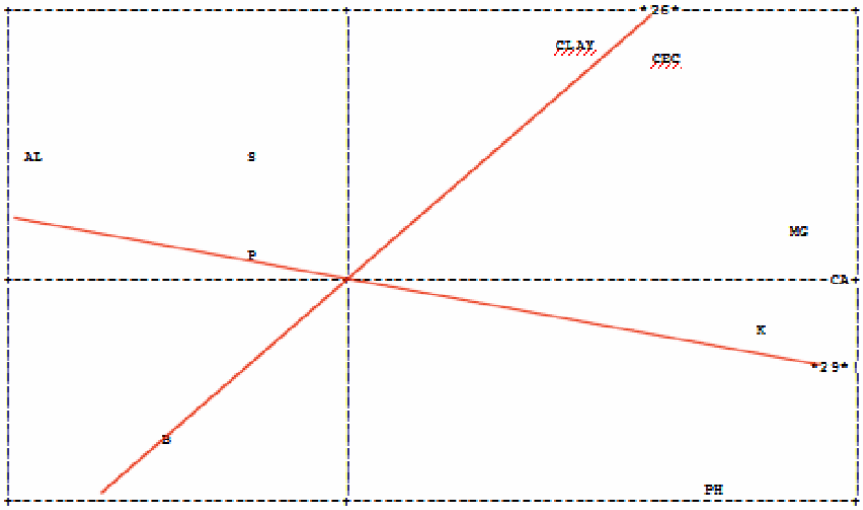


FIG. 1 – *Le plan factoriel associé au noeud 30 de la hiérarchie des variables de sol.*

3.2 Fréquences

Dans cet exemple, on étude le contenu iconographique des images de sceaux Mésopotamiens de la période Uruk-Jamdat Masr. Il s'agit de petits cylindres de pierre dont les images gravées viennent imprimées, lorsqu'on les fait rouler, sur la craie ou sur autre support.

Pour le codage des images, on les a décrites à l'aide d'un texte formalisé qui a été ensuite traité par des analyses textuelles (telle que l'Analyse des Correspondances Textuelles, Lebart et Salem, 1994; Camiz et Rova, 2001). Ici le tableau de contingence, croisant 834 images avec 169 formes textuelles, a été soumis à la classification hiérarchique factorielle basée sur l'*AFC*. Dans le Tableau 2 on voit la succession des deuxièmes valeurs-propres de la hiérarchie, ainsi que l'inertie synthétisée par les variables représentatives des groupes formés.

NUM	*NI*	*NJ*	*2ème v.p.*	*% inertie*	*inertie cumul*	*histogramme*
328	326	276	0.14934	1.41777	71.17208	**********
329	327	313	0.15485	1.47010	72.64218	**********
330	**329**	**328**	**0.16104**	**1.52884**	**74.17102**	**********
331	226	145	0.20507	1.94681	76.11784	**********
332	330	331	0.21873	2.07651	78.19435	**********
333	332	287	0.24908	2.36463	80.55899	**********
334	**333**	**274**	**0.26420**	**2.50822**	**83.06722**	**********
335	**46**	**49**	**0.34822**	**3.30584**	**86.37306**	**********
336	334	335	0.42892	4.07197	90.44504	**********
337	336	303	0.45188	4.28989	94.73494	**********
Total var. representative			0.55460	5.26505	100.00000	

TAB. 2 – *La progression des indices de la hiérarchie des formes lexicales des sceaux.*

Dans la Figure 2 on voit le plan factoriel associé à l'avant-dernier nœud, avec les formes plus intéressantes des deux groupes, ainsi que les images des sceaux les plus significatifs de ce type.

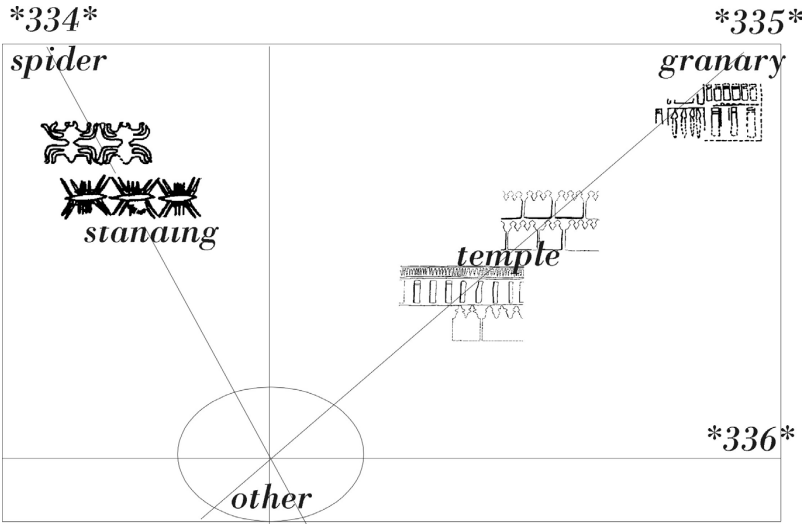


FIG. 2 – *Les formes les plus significatives associées aux variables représentatives des groupes 334 et 335 se fusionnant dans le nœud 336 et les images des sceaux correspondants les plus typiques.*

4 Conclusion

L'objectif des méthodes proposées dans cette nouvelle approche est d'obtenir une information synthétique de manière plus aisée et plus rapide pour l'utilisateur. En particulier on a remarqué que le plan factoriel associé au dernier nœud montre très souvent une structure très proche de celle issue d'une *ACP* où d'une *AFC*. En plus, si l'objectif de l'analyse porte sur l'identification des caractères les plus intéressants pour des études ultérieures, leur sélection, par exemple à l'aide de leur corrélation avec les variables représentatives des groupes, peut s'avérer fort intéressante. De même, si on cherche des règles d'extraction de données, on peut associer aux analyses une segmentation des unités pas à pas (Camiz et Denimal, 2003).

Il reste à comparer ces méthodes avec d'autres méthodes de classification, basées sur d'autres critères, en particulier avec ces de Lerman (1981) où l'idée des facteurs découle aussi des classifications.

Des généralisations semblent également possibles, soit relativement à d'autres types de données (plusieurs caractères qualitatifs, tableaux multiples), soit dans d'autres domaines, tels que l'analyse discriminante, les arbres de décision, etc.

Références

Camiz S. et J.J. Denimal (2003). Nouvelle technique de segmentation associée à une classification de variables. *XXXVèmes Journées de Statistique*, Lyon, 2-6- juin 2003. Société Française de Statistique, Université Lumière Lyon 2, 1: 293-296.

Camiz S. et E. Rova (2001). Exploratory Analyses of Structured Images: a Test on Different Coding Procedures and Analysis Methods. *Archeologia e calcolatori*, 12 : 7-46.

Denimal J.J. (1997). Aides à l'interprétation mutuelle de deux hiérarchies construites sur les lignes et les colonnes d'un tableau de contingence. *Revue de Statistique Appliquée*, 4.

Denimal J.J, (2000). Correspondances hiérarchiques : une nouvelle approche. XXXII Journées de Statistique 15-19 mai 2000 Fes-Maroc.

Denimal J.J. (2001). Hierarchical Factorial Analysis. *Actes du 10th International Symposium on Applied Stochastic Models and Data Analysis*. Compiègne.

Lebart L. et A. Salem (1994). *Statistique textuelle*. Paris : Dunod.

Lerman I.C. (1981). *Classification et analyse ordinale des données*. Paris, Dunod.

Pillar V.D., A.V.A. Jacques et I.I. Boldrini (1992). Fatores de ambiente relacionados à variação da vegetação de um campo natural. *Pesquisa Agropecuária Brasileira,* 27 : 1089-1101.

Qannari E.M., E. Vigneau et Ph. Courcoux (1999). Classification des variables autour de composantes principales; applications». *XXXI Journées de Statistique - 17-21 mai 1999 - Grenoble France*, Résumés, Société Française de Statistique : 677-679.

SAS Institute (1999), *SAS Online Doc*, Version 8, Cary, NC, SAS Institute Inc.

Summary

In this paper, two methods are introduced for ascendant hierarchical classification of variables both quantitative characters and frequencies. They are based, at every step, on a factor analysis of the representative characters of the two merging nodes. Through this method, at any step, a factor plane is built and represents both the characters belonging to the two merging groups and the whole set of the units as seen only by this subset of characters. The examination of these factor planes allows us to obtain an easier interpretation of groups and nodes. Furthermore, a segmentation of the units according to the variables groups can also be defined. The behaviour of the methods is shown on real examples.