ANALYSE FACTORIELLE MULTIPLE DE DONNEES MIXTES : APPLICATION A LA COMPARAISON DE DEUX CODAGES

Jérôme Pagès[1] & Sergio Camiz[2]

[1] Laboratoire de mathématiques appliquées Agrocampus Rennes 65 rue de Saint Brieuc CS 84215 – 35000 RENNES

[2] Università di Roma La Sapienza, Dipartimento di Matematica Guido Castelnuovo, Piazzale Aldo Moro 2, Roma, Italy, I 00185

Résumé.

L'analyse factorielle multiple (AFM) est appliquée à un ensemble de variables d'échelles considérées à la fois comme quantitatives et qualitatives. On illustre ainsi une façon de prendre en compte, dans une analyse factorielle, ces deux types de variables simultanément en tant qu'éléments actifs.

Mots clés. *Analyse factorielle multiple, données mixtes, échelles.*

Summary

The multiple factor analysis (MFA) is applied to a whole of scale variables considered both as quantitative and qualitative variables. One thus illustrates a way of taking into account these two types of variables simultaneously as active in a factor analysis.

Key words. *Multiple factor analysis, mixed data, scale data.*

1 Introduction

Il est souvent souhaité, par les utilisateurs, d'introduire simultanément des variables qualitatives et quantitatives en tant qu'éléments actifs dans une même analyse factorielle. L'analyse factorielle de données mixtes (AFDM ; Pagès 2004), qui reprend les méthodes proposées par B. Escofier (1979a) et G. Saporta (1990), répond bien à cette problématique. Par ailleurs, lorsque l'on dispose de groupes de variables quantitatives et qualitatives, l'analyse factorielle multiple (AFM) est bien appropriée (Escofier & Pagès 1998 ; Pagès, 2002). Pour aider les utilisateurs à mettre en œuvre ces méthodologies, il est nécessaire de diffuser des applications montrant comment concrètement :

- on peut construire des axes factoriels en s'appuyant de façon équilibrée sur les deux types de variables ;
- on peut mener une interprétation en intégrant les deux types de variables.

Pour cela, nous utiliserons une application à caractère méthodologique : la comparaison entre deux façons de prendre en compte des échelles de notation dans les questionnaires. Dans les enquêtes, on demande souvent d'évaluer par une note le degré d'accord à un item, l'importance accordée à tel critère, la satisfaction induite par tel produit au service, etc. On dispose ainsi d'un ensemble de variables qui peuvent être considérées comme quantitatives (ce qui conduit à mettre en œuvre une ACP) ou qualitative (ce qui conduit à une ACM). Chaque point de vue présente ses avantages et inconvénients. L'objectif de cette communication est … multiple :

- comment l'analyse factorielle gère-t-elle des données mixtes actives ?
- en quoi l'AFM est-elle un outil commode pour comparer différents codage d'un même ensemble de variables ?
- faut-il traiter, dans les questionnaires, les échelles de notations en tant que variables quantitatives ou qualitatives ?

2 Notations

Nous disposons de I individus. Chaque individu i est muni du poids p_i tels que $\Sigma_i p_i = 1$. Pour simplifier, nous supposons les individus de même poids $p_i = 1/I \ \forall i$. Ces individus sont décrits par :

- K variables quantitatives $\{v_k \ ; \ k = 1, \ K\}$; ces variables seront toujours supposées centrées réduites ;
- Q variables qualitatives $\{V_q \ ; \ q = 1, \ Q\}$

Ces notations peuvent être rassemblées dans le tableau de la figure 1 dans lequel les variables qualitatives apparaissent à la fois sous leur forme condensée et sous leur forme disjonctive complète.

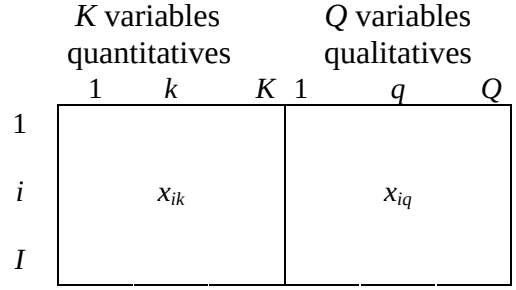

Figure 1. *Structure des données et principales notations.*
x_{ik} : valeur de i pour la variable v_k.
x_{iq}: modalité de i pour la variable V_q.

3 Critères

Dans un premier temps, nous considérons les variables dans leur ensemble, c'est-à-dire non structurées en groupes. On est dans le cadre de l'AFDM. Dans l'espace des variables (R^I), le critère maximisé par l'axe de rang s peut s'écrire :

$$\lambda_s = \sum_{k \in K} r^2(z_s, v_k) + \sum_{q \in Q} \eta^2(z_s, V_q)$$

en notant :

- $r^2(z_s, v_k)$ le carré de leur coefficient de corrélation entre v_k et le facteur z_s de rang s ;
- $\eta^2(z_s, V_q)$ le carré de leur rapport de corrélation entre V_q et le facteur z_s de rang s ;
- λ_s la valeur propre de rang s.

Saporta (1990) semble être le premier à avoir proposé ce critère sous cette forme dans le cadre de l'ACP. Auparavant, Escofier (1979a) avait proposé ce même critère mais sous une forme géométrique et dans le cadre de l'ACM. Soit :

$$\lambda_s = \sum_{k \in K} \cos(\theta_{v_s}^k)^2 + \sum_{q \in Q} \cos(\theta_{v_s}^q)^2$$

en notant

- $\theta_{v_s}^k$: l'angle dans R^I entre z_s et la variable quantitative v_k ;
- $\theta_{v_s}^q$: l'angle dans R^I entre z_s et sa projection sur le sous-espace engendré par les indicatrices de la variable qualitative V_q .

Si l'on prend en compte la structure en groupes des variables qui ici coïncide avec leur type, il convient de mettre en œuvre une AFM. Dans cette analyse, les variables du groupe j sont pondérées par $1/\lambda_1^j$ en notant λ_1^j la première valeur propre de l'analyse séparée du groupe j (une ACP pour des variables quantitatives, une ACM pour des variables qualitatives). Le critère devient :

$$\lambda_s = \frac{1}{\lambda_1^1} \sum_{k \in K} r^2(z_s, v_k) + \frac{1}{\lambda_1^2} \sum_{q \in Q} \eta^2(z_s, V_q)$$

4 Données

Les données proviennent d'une enquête réalisée auprès des étudiants de la faculté d'Economie de Brescia (Italie du Nord). D'un long questionnaire, nous avons extrait 10 variables concernant :

- leur évaluation de six aspects de l'université ; les espaces pour les cours (e1), les horaires (e2), les espaces pour l'étude (e3), les aides à l'étude (e4 : bibliothèque, centre de calcul, etc.), l'administration en général (e5) et leurs rapports avec les enseignants (e6) ;
- leur satisfaction sur quatre aspects personnels ; leur situation économique (s1), leur situation familiale (s2), leurs rapports amicaux (s3) et l'environnement universitaire (s4).

Pour chaque variable, on demandait de répondre par une note sur une échelle allant de 1 (évaluation très négative/forte insatisfaction) à 9 (évaluation très positive/forte satisfaction).
Finalement, on dispose de 2403 questionnaires et de dix variables. Le nombre de non-réponses varie de 205 à 317 selon la question. Les non-réponses sont codées par une modalité *ad hoc* dans le point de vue ACM et sont remplacées par la moyenne de la variable dans le point de vue ACP. Ces données ont déjà fait l'objet de plusieurs analyses (Camiz et al, 1994 ; Camiz et Tagliacozzo, 1997).

5 Analyses séparées

		axe 1	axe 2	axe 3	axe 4	axe 5	axe 6	axe 7	axe 8	axe 9	axe 10
Valeur	ACP	0,666	0,341	0,279	0,234	0,195	0,178	0,164	0,149	0,143	0,136
propre	ACM	3,280	1,695	0,857	0,775	0,730	0,638	0,608	0,543	0,493	0,381
%	ACP	7,401	3,794	3,101	2,595	2,170	1,977	1,827	1,656	1,585	1,514
	ACM	32,797	16,952	8,568	7,749	7,304	6,385	6,079	5,430	4,931	3,806
%	ACP	7,401	11,195	14,296	16,891	19,061	21,038	22,865	24,521	26,105	27,619
cumul	ACM	32,797	49,749	58,317	66,066	73,369	79,754	85,833	91,263	96,194	100,00

Tableau 1. *Inerties des analyses séparées.*

L'ACP met en évidence (Tableau 1) deux facteurs sensiblement plus importants que les autres. L'ACM met en évidence d'abord un facteur bien marqué puis trois autres après lesquels la décroissance des inerties est très régulière. Ces allures sont assez typiques des décroissances de valeurs propres que l'on observe en pratique en ACP et ACM. Dans cette différence d'allure réside un difficulté majeure pour l'analyse simultanée des deux types de variables : la pondération de l'AFM neutralise bien la différence au niveau de la première valeur propre mais, évidemment, ne va pas au delà (ce qui ne serait pas souhaitable puisque pour cela il faudrait modifier la structure interne des sous-tableaux).

		Axes de l'ACM									
		axe 1	axe 2	axe 3	axe 4	axe 5	axe 6	axe 7	axe 8	axe 9	axe 10
	axe 1	0,085	0,940	0,315	0,019	0,014	0,028	0,004	0,031	-0,006	0,047
	axe 2	0,014	-0,080	0,220	0,182	-0,183	0,754	0,429	-0,144	0,072	-0,119
	axe 3	-0,023	-0,014	0,063	-0,045	-0,102	-0,072	0,016	-0,010	-0,072	-0,007
Axes	axe 4	-0,020	0,012	-0,015	-0,050	-0,029	0,043	-0,054	-0,038	0,012	0,016
de	axe 5	-0,008	-0,008	0,029	-0,035	0,008	-0,002	-0,012	-0,020	-0,041	0,017
l'ACP	axe 6	0,012	0,000	-0,018	0,009	0,050	0,011	0,055	0,003	0,010	0,008
	axe 7	-0,009	-0,027	0,081	-0,038	-0,063	-0,049	-0,031	-0,040	0,125	-0,018
	axe 8	-0,025	-0,006	0,028	-0,067	-0,037	-0,026	0,050	-0,012	0,036	0,039
	axe 9	0,018	0,013	-0,032	0,014	0,041	0,031	-0,052	-0,021	0,037	-0,029
	axe 10	-0,005	0,010	-0,030	0,021	-0,009	-0,011	0,036	0,044	0,077	0,032

Tableau 2. *Corrélations entre facteurs des analyses séparées*

La matrice des corrélations (Tableau 2) met en évidence :

- une forte liaison entre le premier facteur de l'ACP et le second de l'ACM ;
- une liaison entre le second facteur de l'ACP et les facteurs 6 et 7 de l'ACM.

La structure commune aux deux tableaux ne risque donc pas d'apparaître en comparant simplement les résultats des deux analyses. Une méthode d'analyse globale est nécessaire.

6 Analyse Factorielle Multiple

6.1 Inerties (Tableau 3)

Les valeurs propres de l'AFM mettent en évidence trois facteurs bien marqués. Leur décomposition par groupe montre que le premier et le troisième facteur correspondent à des directions d'inertie importante des deux groupes, le deuxième étant spécifique du groupe ACM.

	axe 1	axe 2	axe 3	axe 4	axe 5	axe 6	axe 7	axe 8	axe 9	axe 10
Ensemble en %	9,107	6,002	4,758	2,612	2,366	2,232	2,083	2,024	1,879	1,793
Ensemble	1,508	0,994	0,788	0,433	0,392	0,370	0,345	0,335	0,311	0,297
Groupe 1 (ACM)	0,516	0,985	0,280	0,340	0,218	0,186	0,221	0,185	0,170	0,157
Groupe 2 (ACP)	0,992	0,009	0,508	0,093	0,174	0,184	0,124	0,150	0,141	0,140

Tableau 3. *Inertie des axes de l'AFM, globale et ventilée par groupe*

6.2 Représentations des variables quantitatives et des modalités (Figures 2 et 3)

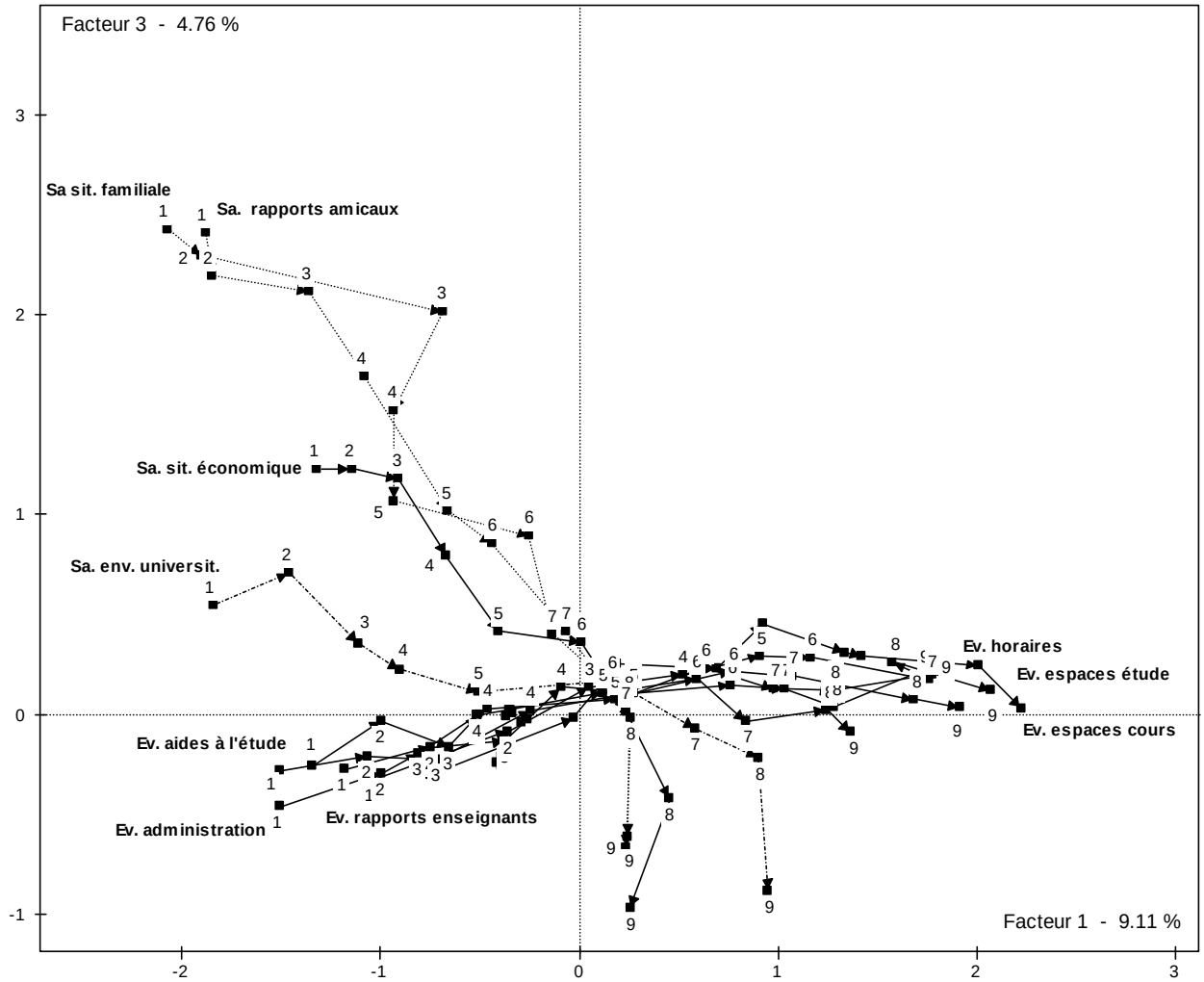

Figure 2. Représentation des modalités sur le plan 1-3 de l'AFM

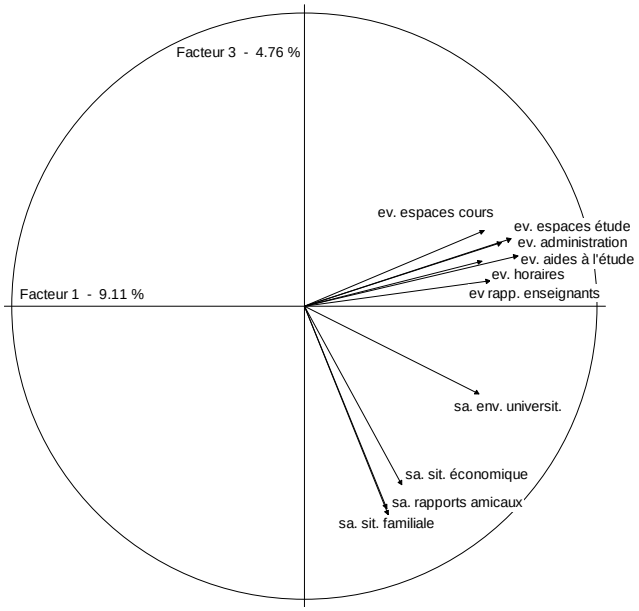

Figure 3. Cercle des corrélations.

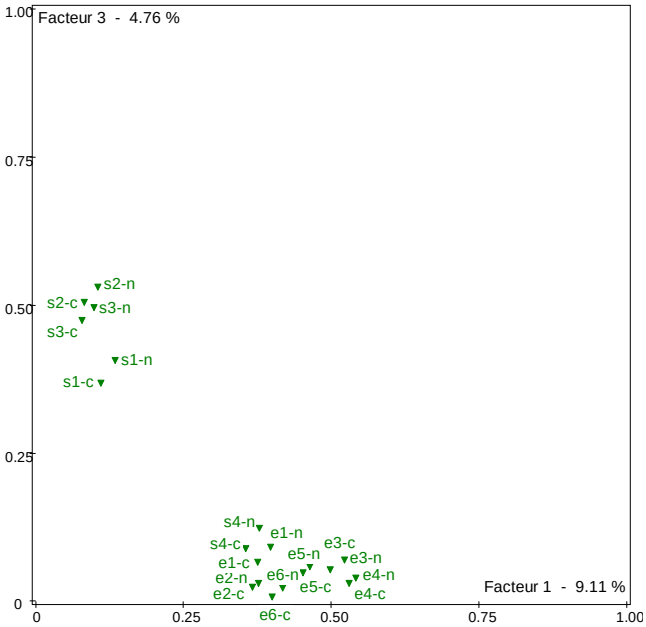

Figure 4. Carré des liaisons.

En AFM sur données mixtes, comme en AFDM, chaque modalité est représentée par le centre de gravité (exact et non à $1/\lambda_s$ près comme en ACM) des individus qui la possèdent. Les variables quantitatives sont représentées comme en ACP.

Le second facteur (non représenté), spécifique du groupe ACM, classe les individus selon leur nombre de non-réponses. Un tel facteur ne peut évidemment pas apparaître via les variables quantitatives. Nous examinons ci-après plutôt le plan 1-3 commun aux deux groupes.

Le premier axe est corrélé à toutes les variables : il oppose les individus ayant une évaluation positive pour les 6 aspects de l'université et étant satisfait des 4 aspects de leur vie personnelle, aux individus ayant émis les opinions opposées. Le troisième axe met en évidence la spécificité de la situation personnelle vis à vis de l'évolution de l'université. Remarquons, au passage, le caractère intermédiaire de la satisfaction quant à l'environnement universitaire.

Dans cette présentation méthodologique, nous ne détaillons pas plus le commentaire de ce graphique.

6.3 Représentation simultanée des variables des deux types (figure 4)

Comme en AFDM, on peut représenter sur un même graphique :

- les variables quantitatives par le carré de leur coefficient de corrélation avec les facteurs ;
- les variables qualitatives par le carré de leur rapport de corrélation avec les facteurs.

Ce graphique est dit « carré des liaisons ». Il peut s'interpréter comme une projection (Escofier & Pagès 1998). Il trouve son origine dans les représentations de variables qualitatives proposées par Escofier (1979b) et Cazes (1982) dans le cadre de l'ACM.

La figure 4 fait clairement apparaître deux groupes de variables : les six évaluations, surtout liées au facteur 1et trois satisfactions, liées surtout au facteur 3. La position de la satisfaction vis à vis de l'environnement universitaire, clairement du côté des évaluations.

Cette figure donne une image des liaisons variables-facteurs plus tranchée que les graphiques des figures 2 et 3. Cela tient à l'élévation au carré du coefficient ou du rapport de corrélation. Ce graphique sera donc précieux pour un débroussaillage parmi de très nombreuses variables.

Sur la figure 4, les images quantitative et qualitative d'une même variable sont très proches. Cela tient à la linéarité des liaisons mises en évidence sur ce plan. Ces points seraient évidemment éloignés si l'on considérait l'axe 2.

7 Conclusion

Quelques éléments de réponse aux trois questions qui terminent l'introduction.
La pondération de l'AFM permet d'équilibrer les deux types de variables dans la construction des axes. Pour les variables, l'AFM ;présentée fournit :

- Les représentations usuelles de l'ACP et de l'ACM.
- Une représentation supplémentaire qui intègre les deux types de variables.

L'AFM met en évidence des structures communes aux groupes de variables qui n'apparaissent pas forcement dans une comparaison directe des facteurs des analyses séparées. Les structures spécifiques des groupes apparaissent aussi sur les premiers axes lorsqu'elles sont importantes.
Dans le traitement d'une batterie d'échelles, ACP et ACM sont complémentaires, fournissant chacune son compromis entre simplicité et richesse. L'AFM peut aussi être utilisée pour bénéficier des deux points de vue dans une analyse en quelque sorte « robuste » vis à vis du codage.

Bibliographie

[1] CAMIZ, S., M. CIVARDI, R. CREMONESI, P. FALBO, S. STEFANI ET S. ZOPPETTI (1994), « *Indagine sullapopolazione studentesca dell'Università di Brescia* », Rapport de recherche du Dipartimento di Metodi Quantitativi dell'Università di Brescia, Quaderno n. 62.

[2] CAMIZ, S. , ET G. TAGLIACOZZO (1997), « Analisi testuale delle risposte a testo libero dei questionari degli studenti di Brescia: primi risultati », dans M. Bottiroli Civardi et S. Camiz (eds.), *La popolazione studentesca e le Università Italiane: indagini, modelli e risultati*. Padova : CLEUP Editrice, p. 211-241.

[3] CAZES, P. (1982). Note sur les éléments supplémentaires en analyse des correspondances. *Les cahiers de l'Analyse des données,* 7 (1), p. 9-23 et 7 (2), p. 133-154.

[4] ESCOFIER, B. (1979a). Traitement simultané de variables quantitatives et qualitatives en analyse factorielle. *Les cahiers de l'analyse des données* 4 (2) 137-146.

[5] ESCOFIER, B. (1979b). Une représentation des variables dans l'analyse des correspondances multiples. *Revue Statistique Appliquée* XXVII (4) 37-47.

[6] ESCOFIER, B. ET PAGES, J. (1998). *Analyses factorielles simples et multiples*. 3e ed. Dunod.

[7] FACTOMINER (2007). Logiciel libre d'analyse factorielle en R, diffusé par le laboratoire de mathématiques appliquées d'Agrocampus. http://factominer.free.fr/

[8] PAGES, J (2002). Analyse factorielle multiple appliquée aux variables qualitatives et aux données mixtes. *Revue de statistique appliquée* L (4) 5-37.

[9] PAGES, J.(2004). Analyse Factorielle de Données Mixtes. *Revues de Statistique Appliquée,* LII (4), 93-111.

[10] SAPORTA, G. (1990). *Probabilités, analyse des données et statistique*. Technip, Paris.