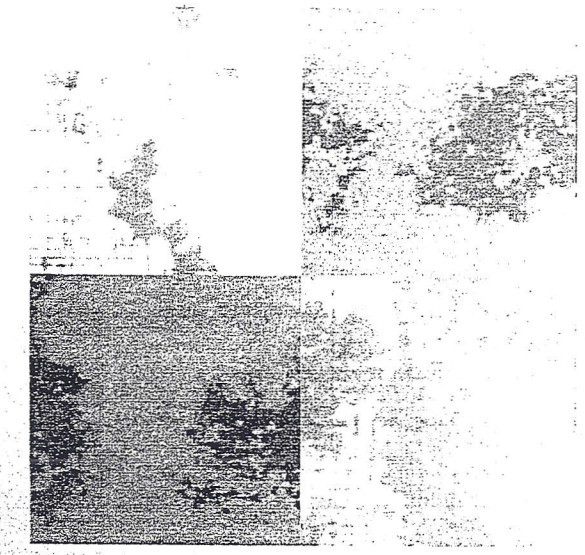
www.cepadues.com

ISSN 1764-1667

Réf. 831
ISBN : 978.2.85428.831.5

9 782854 288315

Revue
Nouvelles
Technologies
Information



Classification : points de vue croisés

Rédacteurs invités :
Mohamed Nadif et François-Xavier Jollois

RNTI-C-2

Cépaduès
éditions

Analyse en Composantes Principales Mixte

Rafik Abdesselam

CREM UMR CNRS 6211
Université de Caen, Esplanade de la Paix, F-14032 Caen, France
rafik.abdesselam@unicaen.fr,
http://www.unicaen.fr/crem/membres.php

Résumé. Le traitement simultané de données mixtes (quantitatives et qualitatives) ne peut pas se réaliser directement par les méthodes classiques de la statistique exploratoire multidimensionnelle. Dans ce travail, l'analyse factorielle sur données mixtes proposée est une analyse en composantes principales normée après transformation des indicatrices des variables qualitatives en variables quantitatives au travers de projections de nuages de points dans l'espace des individus correspondant à des analyses de la variance multivariée. La méthode est évaluée sur la base d'une application sur données réelles mixtes.

1 Introduction

Dans le cadre d'un traitement de données multidimensionnelles, il est très fréquent que le thème homogène de variables à analyser soit constitué de données mixtes (variables quantitatives et qualitatives). La méthodologie usuelle de traitement consiste, soit à mettre les variables qualitatives [resp. quantitatives] en éléments illustratifs dans une Analyse en Composantes Principales (ACP) [resp. Analyse des Correspondances Multiples (ACM)], soit encore de discrétiser les variables quantitatives du thème en variables qualitatives en vue d'une ACM, ce qui introduit très souvent un biais dû au choix du nombre de classes et de leurs amplitudes égales ou différentes, et qui occasionne une perte d'information. De nombreux chercheurs se sont intéressés à cette problématique et ont proposé des méthodes qui traitent simultanément les deux types de variables en éléments actifs : l'ACP avec indicatrices introduite par Tenenhaus (1977); Escofier et Pagès (1979); Saporta (1990); Young (1981); Pagès (2002) et plus récemment l'Analyse Factorielle de Données Mixtes (AFDM) proposée par Pagès (2004).

L'Analyse en Composantes Principales Mixte (ACPM) proposée est une ACP normée des variables quantitatives et des indicatrices des variables qualitatives transformées en variables combinaisons linéaires des variables quantitatives, à partir de projections orthogonales de nuages de points dans l'espace des individus muni d'un produit scalaire relationnel. Chaque variable qualitative est quantifiée séparément, puis traitée de la même façon que les variables quantitatives.

L'ACPM est formellement proche de l'AFDM en ce sens qu'elles consistent toutes les deux à quantifier les variables qualitatives en vue d'une ACP, mais elles se différencient toutefois par le type d'ACP et le choix de la tranformation.

2 Transformation des données qualitatives en quantitatives

On dispose de p variables quantitatives centrées $\{x^j; j = 1, p\}$ et de m variables qualitatives $(y_1, \ldots, y_l, \ldots, y_m)$ auxquelles sont associées au total $q = \sum_{l=1}^{m} q_l$ variables indicatrices non centrées $\{y_l^k; k = 1, q_l\}_{\{l=1,m\}}$.

On utilisera les notations suivantes pour construire la matrice M associée au produit scalaire de référence dans l'espace des individus $E = E_x \oplus E_y = R^{p+q}$.

$E_x = R^p$ étant le sous-espace des individus associé par dualité aux p variables quantitatives centrées $\{x^j; j = 1, p\}$,
$E_y = \oplus\{E_{y_l}\}_{\{l=1,m\}} = R^q$ étant les m sous-espaces des individus associés aux m variables qualitatives $(y_1, \ldots, y_l, \ldots, y_m)$,
$E_{y_l} = R^{q_l}$ étant le sous-espace des individus associé par dualité aux q_l variables indicatrices non centrées des modalités de y_l,
X est la matrice d'ordre (n, p) des données quantitatives associée aux p variables centrées $\{x^j; j = 1, p\}$,
Y_l est la matrice des données qualitatives d'ordre (n, q_l) associée aux q_l indicatrices non centrées $\{y_l^k; k = 1, q_l\}$ de la $l^{ème}$ variable qualitative y_l, codage disjonctif complet,
M_{y_l} [resp. M_x] est la matrice du produit scalaire dans l'espace E_{y_l} [resp. E_x], isomorphe du sous-espace de même nom, via l'injection canonique,
D_{y_l} est la matrice diagonale des poids définie par $[D_{y_l}]_{kk} = n_k/n$ pour tout $k = 1$ à q_l, où n_k est le nombre d'individus possédant la modalité k de y_l,
$N_x = \{x_i \in E_x\ ;\ i = 1, n\}$ est le nuage des individus associé au tableau de données X,
$N_{y_l} = \{y_i \in E_{y_l}\ ;\ i = 1, n\}$ est le nuage des individus associé au tableau de données Y_l,
$P^M_{E_{y_l}}$ est l'opérateur de projection M-orthogonale sur E_{y_l}.

La transformation des données qualitatives en quantitatives se fait à l'aide de la construction statistique et géométrique de m nuages $\hat{N}_x^{y_l} = \{P^M_{E_{y_l}}(x_i); x_i \in N_x\} \subset E_{y_l} \subset E$. Pour tout $l = 1$ à m, le sous-espace E_{y_l} est considéré comme sous-espace explicatif sur lequel est projeté M-orthogonalement le nuage N_x des données quantitatives du sous-espace à expliquer E_x.

Le produit scalaire M de référence dans l'espace $E = E_x \oplus \{E_{y_l}\}_{\{l=1,m\}} = R^{p+q}$ des individus joue un rôle fondamental dans notre approche pour réaliser les m projections. A priori, pour tout $l = 1$ à m, le produit scalaire M_{y_l} intra le sous-espace E_{y_l} sur lequel on projette, pourrait être quelconque (Abdesselam (2005)) ; le choix $M_{y_l} = \chi^2_{y_l} = D_{y_l}^{-1}$ (distance du khi-deux) simplifie les calculs. Quant au produit scalaire M_x, intra le sous-espace E_x, on choisit $M_x = V_x^+$ (distance de Mahalanobis) afin de maximiser le critère d'inertie expliquée.

La matrice M d'ordre $(p + q)$ associée au produit scalaire partitionné et équilibré dans E, relativement à l'ensemble des couples de variables $\{x^j; j = 1, p\}$ et $\{y_l{}^k; k = 1, q_l\}_{l=1,m}$, est telle que :

$$\begin{cases} M_x = V_x^+ \quad ; \quad M_{y_l} = \chi^2_{y_l} & pour\ l = 1, m \\ M_{xy_l} = M_x[(V_x M_x)]^+ V_{xy_l} M_{y_l}[(V_{y_l} M_{y_l})]^+ = V_x^+ V_{xy_l} \chi^2_{y_l} & pour\ l = 1, m \\ M_{y_l y_{l'}} = 0 & pour\ l \neq l' \end{cases}$$

où $V_{y_l} = {}^tY_l D Y_l$, $V_x = {}^tXDX$ et $V_{xy_l} = {}^tXDY_l$ désignent les matrices de variances-covariances, $D = (1/n)I_n$ est la matrice diagonale des poids des n individus et I_n la matrice unité d'ordre n. $[(V_x M_x)]^+]$ est l'inverse généralisée de Moore-Penrose de $(V_x M_x)]$. L'introduction d'inverses généralisées est une conséquence de la singularité des matrices V_x et V_{y_l}, lorsque rang $[V_x] < p$ et puisque rang $[V_{y_l}] = q_l - 1$.

Le produit scalaire M positionne les sous-espaces des individus E_x et $\{E_{y_l}\}_{\{l=1,m\}}$ tel que l'on puisse traduire en terme d'inertie dans l'espace des individus E, la structure des associations observées entre les sous-espaces des variables F_x et F_y associés par dualité dans l'espace des variables $F = R^n$ muni de la métrique diagonale des poids D.

Par ailleurs, il n'est pas indispensable de travailler dans l'espace E des individus, on peut tout aussi bien raisonner dans l'espace F des variables.

Par construction, le produit scalaire dans l'espace explicatif $E_y = \oplus\{E_{y_l}\}_{\{l=1,m\}}$ est à effet relationnel nul. Ainsi, la projection M-orthogonale du nuage N_x sur E_y a pour coordonnées $[\hat{X}^{y_1}, \ldots, \hat{X}^{y_l}, \ldots, \hat{X}^{y_m}]$: juxtaposition de m tableaux $\hat{X}^{y_l} = XV_x^+ V_{xy_l}$ de données quantitatives d'ordre (n, q_l) associés aux m nuages de points $\hat{N}_x^{y_l} = P^M_{E_{y_l}}(N_x) \subset E_{y_l}$. On obtient ainsi l'ensemble des m projections : c'est là un des avantages pratiques du produit scalaire M dans l'espace E des individus.

Remarque : L'ACP du triplet $(\hat{X}^{y_l}; \chi^2_{y_l}; D)$ est équivalente au MANOVA : l'analyse de la variance multivariée entre les p variables quantitatives et les q_l indicatrices associées aux niveaux du facteur explicatif y_l, et dont l'inertie expliquée, $I(\hat{N}_x^{y_l}) = trace(V_{y_l x} V_x^+ V_{xy_l} \chi^2_{y_l})$, est égale à la trace de Pillai.

On note, pour tout $l = 1$ à m, $Z_l = Y_l - {}^tG_l$ la matrice des données d'ordre (n, q_l) associée au nuage $N_{z_l} \subset E_{y_l}$ dont l'inertie $I(N_{z_l}) = q_l - 1$. $G_l = {}^t\hat{X}^{y_l} D\, 1_n$ désigne le vecteur moyennes des variables de $\hat{X}^{y_l}$ et 1_n le vecteur unité d'ordre n.

Définition : L'ACPM du tableau de données mixtes $[X, Y_1, \ldots, Y_l, \ldots, Y_m]$ d'ordre $(n; p+q)$ consiste à effectuer l'ACP normée du tableau $[X, Z_1, \ldots, Z_l, \ldots, Z_m]$ de données (quantitatives).

Ainsi, les m tableaux d'indicatrices non centrées $[Y_1, \ldots, Y_l, \ldots, Y_m]$ associés aux m variables qualitatives sont centrés relativement aux m centres de gravité $[G_1, \ldots, G_l, \ldots, G_m]$, puis remplacés par les m tableaux $[Z_1, \ldots, Z_l, \ldots, Z_m]$ de variables quantitatives via les m tableaux $[\hat{X}^{y_1}, \ldots, \hat{X}^{y_l}, \ldots, \hat{X}^{y_m}]$ des m MANOVA séparées.

D'un point de vue méthodologique, l'ACPM se présente comme un enchaînement de deux procédures : une procédure de projection de nuages de points correspondant aux coordonnées de MANOVA afin de quantifier les données qualitatives, on tient ainsi compte des rapports de corrélation, puis une procédure d'ACP normée pour synthétiser les corrélations linéaires de l'ensemble des variables quantitatives et qualitatives transformées.

D'un point de vue pratique, il suffit simplement de centrer chaque tableau d'indicatrices Y_l par rapport au centre de gravité correspondant $G_l = V_{y_l x} V_x^+ {}^t X D 1_n$, puis d'exécuter un programme classique d'ACP standardisée.

Cette analyse mixte réduit le nombre de dimensions des données passant d'un groupe de $(p+q)$ variables initiales à un groupe plus petit dont l'inertie totale est égale et synthétisée par $(p+q-m)$ composantes non corrélées : somme des inerties de l'ACP normée des variables quantitatives et de l'ACM des variables qualitatives.

2.1 Comparaison

L'ACPM coïncide avec l'AFDM proposée dans Pagès (2004) qui consiste à effectuer l'ACP usuelle du tableau de données $[X^{cr}, Y_1/\sqrt{D_{y_1}}, \ldots, Y_l/\sqrt{D_{y_l}}, \ldots, Y_m/\sqrt{D_{y_m}}]$: les variables quantitatives X^{cr} sont centrées et réduites, et les indicatrices des variables qualitatives sont affectées d'une pondération. Pour tout $l = 1$ à m, cela revient à diviser les valeurs de l'indicatrice y_l^k de la variable y_l par $\sqrt{\frac{n_k}{n}}$: codage ACP de l'indicatrice y_l^k.

L'objectif de l'ACPM est le même que celui de l'AFDM, à savoir rechercher les facteurs principaux, notés F^s, qui maximisent le critère mixte ci-dessous, proposé en termes de carrés de corrélations par Saporta (1990) et géométriquement en termes de cosinus carrés d'angles par Escofier (1979) :

$$\sum_{j=1}^{p} r^2(x^j, F^s) + \sum_{l=1}^{m} \eta^2(y_l, F^s) = \sum_{j=1}^{p} \cos^2 \theta_{js} + \sum_{l=1}^{m} \cos^2 \theta_{ls}$$

où r^2 et η^2 sont respectivement le carré du coefficient de corrélation linéaire des variables quantitatives et le rapport de corrélation des variables qualitatives avec le facteur de rang s, et θ l'angle entre les vecteurs correspondants, les variables étant centrées et réduites.

3 Exemple d'application

Pour illustrer cette approche, nous reprenons les données publiées dans Lambin (1990) et reprises en annexe (tableau 7), elles portent sur un échantillon de 27 petites voitures du marché belge. On dispose d'un thème homogène de 9 variables mixtes dont $p = 6$ caractétristiques continues : la cylindrée, la consommation urbaine, la vitesse maximum, le volume du coffre, le rapport poids/puissance et la longueur, et $m = 3$ caractéristiques nominales : la puissance fiscale (4CV, 5CV, 6CV), la marque du constructeur (Française, Etrangère) et quatre classes de prix (CP1, CP2, CP3, CP4) totalisant $q = 9$ modalités.

L'objectif est de synthétiser simultanément au sens des corrélations l'ensemble de ces caractéristiques mixtes.

Libellé	Moyenne	E. Type	Min	Max	FISC	MARQ	PRIX
CONS	7.14	1.12	5.60	9.30	0.809	0.009	0.636
CYLI	1165.63	204.17	903.00	1597.00	0.843	0.002	0.846
VITE	154.26	21.94	115.00	200.00	0.690	0.010	0.826
VOLU	901.41	301.67	202.00	1200.00	0.136	0.168	0.029
RP/P	18.65	5.42	10.20	33.10	0.562	0.042	0.660
LONG	3.62	0.07	3.40	3.70	0.094	0.022	0.163

TAB. 1 – *Statistiques sommaires - Rapports de corrélation.*

Le tableau 1 récapitule les statistiques élémentaires des variables quantitatives ainsi que leurs rapports de corrélation avec les variables qualitatives considérées.

G_1			G_2		G_3			
4CV	5CV	6CV	FRAN	ETRA	CP1	CP2	CP3	CP4
-4.386	1.532	2.853	6.486	-6.487	-8.599	4.236	0.523	3.839

TAB. 2 – *Moyennes pour le centrage des indicatrices.*

Le tableau 2 présente les moyennes ayant servi à la transformation des indicatrices des trois variables qualitatives (puissance fiscale, marque du constructeur et prix) en variables quantitatives, étape préalable à l'application de l'ACP normée de l'ensemble des variables. Les données qualitatives transformées sont reprises en annexe (tableau 8).

	CONS	CYLI	VITE	VOLU	RP/P	LONG	4CV	5CV	6CV	FRAN	ETRA	CP1	CP2	CP3	CP4
CONS	1	0.797	0.780	0.295	-0.682	0.197	-0.548	-0.382	0.896	-0.094	0.094	-0.402	-0.390	0.247	0.656
CYLI		1	0.832	0.112	-0.779	0.290	-0.838	0.061	0.837	0.043	-0.043	-0.677	-0.290	0.481	0.619
VITE			1	0.022	-0.938	0.155	-0.660	-0.145	0.819	-0.100	0.100	-0.603	-0.297	0.325	0.727
VOLU				1	0.102	-0.073	-0.105	-0.266	0.330	0.409	-0.409	-0.011	-0.101	-0.023	0.156
RP/P					1	-0.098	0.609	0.108	-0.735	0.204	-0.204	0.547	0.309	-0.390	-0.580
LONG						1	-0.306	0.165	0.188	0.149	-0.149	-0.402	0.125	0.231	0.113
4CV							1	-0.459	-0.681	-0.125	0.125	0.796	0.113	-0.625	-0.402
5CV								1	-0.337	0.227	-0.227	-0.366	0.264	0.317	-0.199
6CV									1	-0.054	0.054	-0.542	-0.337	0.401	0.590
FRAN										1	-1.000	-0.112	0.227	-0.330	0.328
ETRA											1	0.112	-0.227	0.330	-0.328
CP1												1	-0.366	-0.498	-0.320
CP2													1	-0.309	-0.199
CP3														1	-0.271
CP4															1

TAB. 3 – *Matrice de corrélation des caractéristiques des voitures.*

Les principaux résultats de l'ACPM sont présentés dans les tableaux et graphiques ci-après, ils s'interprètent avec les règles classiques d'une ACP.

Le calcul de la matrice de corrélation des variables, présenté dans le tableau 3, donne des indications sur l'évolution simultanée des variables mixtes prises deux à deux.

	F^1	F^2	F^3	F^4	F^5	F^6	F^7	...	F^{12}
Val.Propre	6.154	2.756	2.349	1.210	0.937	0.808	0.308	...	0.023
% Inertie	41.02	18.37	15.66	8.06	6.25	5.39	2.05	...	0.15
% Cumulé	41.02	59.40	75.06	83.12	89.37	94.75	96.81	...	100.00

TAB. 4 – *Valeurs propres issues de l'ACPM.*

De même, le tableau 4 donne les valeurs propres de la matrice de corrélation, les proportions et les proportions cumulées de la variance expliquée par les composantes. Bien que l'inertie totale soit égale au nombre de variables $p + q = 15$ (ACP sur matrice de corrélation ou normée), elle est résumée et synthétisée par $p + q - m = 12$ moments principaux non nuls.

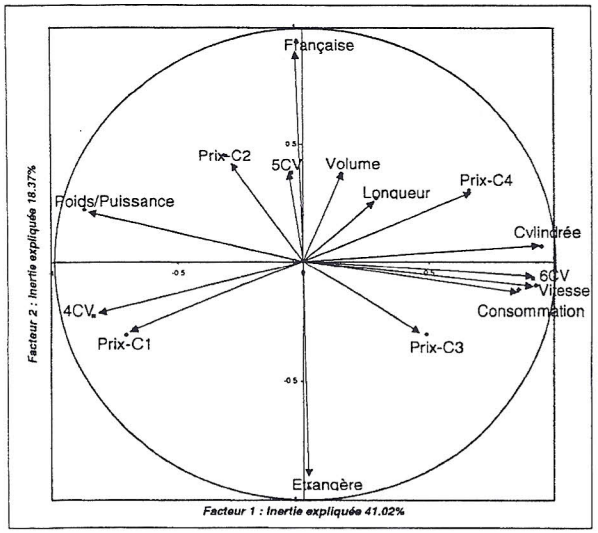


FIG. 1 – *Représentation des variables mixtes dans le premier plan principal.*

La figure 1 donne la représentation géométrique, cercle des corrélations, des variables quantitatives et qualitatives transformées. On peut ainsi interpréter simultanément et en éléments actifs les deux types de variables dans ce nouveau système de coordonnées.

Ainsi, par exemple, l'examen de la figure 1 fait apparaître une opposition entre la cylindrée, la vitesse, la consommation, un prix et une puissance fiscale élevés d'une part (corrélations positives importantes avec le premier axe) et le rapport poids/puissance, une puissance fiscale et un prix faibles d'autre part (corrélations négatives importantes avec le premier axe) ; l'axe 1 représente le degré de performance des voitures auquel sont associés le prix et la puissance fiscale. L'axe 2 est lié à la variable marque du constructeur, il oppose les voitures de marque française aux autres de marque étrangère.

	F^1	F^2	F^3	F^4	F^5	F^6	F^7	...	F^{12}
CONS	**.859**	-.118	.346	.036	.167	.038	.190	...	-.004
CYLI	**.950**	.068	-.059	.030	-.073	.039	.074	...	.020
VITE	**.928**	-.101	.103	-.226	-.101	-.049	-.122	...	.118
VOLU	.152	.375	.477	**.618**	.280	-.275	.081	...	-.003
RP/P	**-.865**	.223	-.010	.235	.167	.062	.254	...	.065
LONG	.292	.270	-.335	-.070	**.506**	**.675**	.004	...	.001
4CV	**-.830**	-.226	.400	-.170	.140	.072	-.154	...	-.004
5CV	-.044	.378	**-.791**	.069	-.401	.007	.212	...	.006
6CV	**.916**	-.072	.228	.123	.182	-.082	-.012	...	-.001
FRAN	-.023	**.946**	.195	.096	-.140	.079	-.155	...	.005
ETRA	.023	**-.946**	-.195	-.096	.140	-.079	.155	...	-.005
CP1	**-.702**	-.302	**.505**	.155	-.221	.256	.052	...	.038
CP2	-.307	.451	-.362	-.455	.436	-.411	.013	...	.015
CP3	.490	-.299	**-.593**	**.520**	.018	.008	-.189	...	-.014
CP4	**.661**	.303	.472	-.381	-.199	.092	.158	...	-.051

TAB. 5 – *Corrélations variables - facteurs.*

Pour évaluer la qualité de la représentation des variables mixtes, le tableau 5 résume les corrélations des facteurs principaux avec les variables initiales.

	F^1	F^2	F^3	F^4	F^5	F^6	F^7	...	F^{12}	Somme
CONS	**.738**	.014	.120	.001	.028	.001	.036	...	.000	1
CYLI	**.903**	.005	.004	.001	.005	.002	.006	...	.000	1
VITE	**.862**	.010	.011	.051	.010	.002	.015	...	.014	1
VOLU	.023	.141	.227	.382	.078	.076	.007	...	.000	1
RP/P	**.748**	.050	.000	.055	.028	.004	.065	...	.004	1
LONG	.085	.073	.112	.005	.256	.456	.000	...	.000	1
FISC	**.918**	.146	**.628**	.029	.163	.007	.049	...	.000	2
MARQ	.001	**.895**	.038	.009	.020	.006	.024	...	.000	1
PRIX	**.929**	.364	**.705**	**.498**	.219	.186	.048	...	.003	3
Somme	5.206	1.697	1.844	1.032	.808	.741	.249	...	.022	12

TAB. 6 – *Résultats du critère.*

Outre les outils d'aide à l'interprétation d'une ACP, les résultats graphiques et numériques, le tableau 6 donne les carrés des corrélations linéaires des variables quantitatives et les rapports de corrélation des variables qualitatives avec les premières composantes principales de l'ACPM.

Si on prend en considération l'ensemble des composantes, la somme des carrés des corrélations étant égale à l'unité pour une variable quantititive, et la somme des rapports de corrélation étant égale à l'inertie totale du nuage associé à la variable qualitative considérée, ce tableau illustre la qualité de la représentation des variables quantitatives et qualitatives sur les axes principaux selon le critère mixte à maximiser.

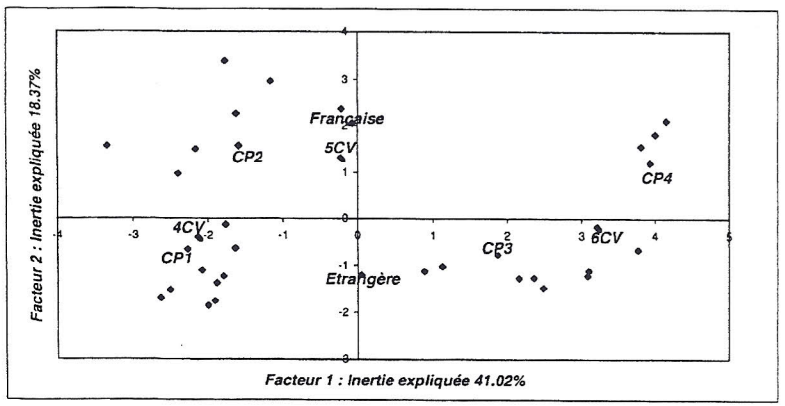


FIG. 2 – *Représentation des individus et des modalités.*

Enfin, la figure 2 donne la représentation graphique classique, dans le premier plan factoriel défini par l'ensemble des variables mixtes initiales, des individus-voitures et des centres de gravité associés aux modalités des variables qualitatives.

4 Conclusion et perspectives

Dans ce travail, l'ACPM proposée semble bien prendre en compte l'équilibre des structures des deux types de variables : les corrélations linéaires entre les variables quantitatives, les associations entre les modalités des variables qualitatives ainsi que leurs rapports de corrélation.

Cette analyse factorielle mixte est facile à mettre en oeuvre à l'aide d'un simple programme d'ACP. Les résultats obtenus sont identiques à ceux de l'AFDM.

Il serait aussi intéressant de comparer les résultats de l'ACPM avec ceux d'une analyse canonique généralisée à $(p+m)$ groupes de variables mixtes, en considérant chaque variable, quantitative ou qualitative, comme un groupe d'une seule variable.

Enfin, cette méthodologie de quantification de variables qualitatives permet d'élargir le champ d'application des techniques de classification aux données mixtes et d'étendre celui des méthodes factorielles, notamment décisionnelles, en mettant en oeuvre une analyse discriminante sur variables mixtes.

Références

Abascal-Fernandez, E., M.-I. Landaluce-Cluo, et I. Garcia-Laube (2003). Multiple factor analysis of mixed tables : a proposal for analysing problematic metric variables. *Proceeding of CARME meeting, Barcelona.*

Abdesselam, R. (2005). *Dissymmetrical Multivariate ANalysis Of VAriance.* Classification and Data Analysis : In Book of Short Papers, Group of the Italian Statistical Society, Editors S. Zani and A. Cerioli, p. 189-192.

Cazes, P. (1980). *Note sur les éléments supplémentaires en analyse des correspondances.* Cahier de l'analyse des données, 7(1) 9-23 et 7(2) 133-154.

Escofier, B. et J. Pagès (1979). *Traitement simultané de variables quantitatives et qualitatives en analyse factorielle.* Cahier de l'analyse des données, , vol. 4(2), p. 137-146.

Gifi, A. (1990). *Nonlinear multivariate analysis.* Leiden : Department of Data Theory.

Kiers, H.-A.-L. (1988). *Principal components analysis on a mixture of quantitative and qualitative data based on generalized correlation coefficients.* M.G.H. Jansen, W.H. van Schuur (Eds).

Lambin, J. (1990). *La recherche marketing, Analyser - Mesurer - Prévoir.* McGraw-Hill.

Pagès, J. (2002). Analyse factorielle multiple appliquée aus variables qualitatives et aux données mixtes. *Revue de Statistique Appliquée L(4)*, 5–37.

Pagès, J. (2004). Analyse factorielle de données mixtes. *Revue de Statistique Appliquée LII(4)*, 5–37.

Saporta, G. (1979). Pondération optimale de variables qualitatives en analyse des données. *Statistique et Analyse des Données 3*, 19–31.

Saporta, G. (1990). *Simultaneous analysis of qualitative and quantitative data.* Atti XXXV Riunione Scientifica della Societa Italiana di Statistica, p. 63-72.

Tenenhaus, M. (1977). Analyse en composantes principales d'un ensemble de variables nominales ou numériques. *Revue de Statistique Appliquée XXV(2)*, 39–56.

Tenenhaus, M. et F. Young (1985). An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika 50*, 91–119.

Young, F. (1981). Qualitative analysis of qualitative data. *Psychometrika 46(4)*, 357–388.

Annexe

NOM	CONS	CYLI	VITE	VOLU	RP/P	LONG	FISC	MARQ	PRIX
AS2	6.20	998	140	955	23.20	3.40	4CV	ETRA	CP1
CI4	5.60	954	145	1170	19.40	3.50	4CV	FRAN	CP1
PE6	6.70	993	145	1151	20.80	3.61	4CV	FRAN	CP2
FI3	6.30	999	140	1088	21.80	3.64	4CV	ETRA	CP1
FI5	6.20	999	145	968	21.50	3.64	4CV	ETRA	CP2
FI8	8.90	1301	200	968	11.00	3.64	6CV	ETRA	CP4
FID	7.70	1302	165	968	16.00	3.64	6CV	ETRA	CP3
FO1	7.00	1117	137	900	22.70	3.64	4CV	ETRA	CP1
RE7	9.30	1597	180	973	12.00	3.64	6CV	FRAN	CP4
NI1	6.40	988	140	375	17.00	3.64	4CV	ETRA	CP1
OP1	7.20	993	143	845	22.40	3.62	4CV	ETRA	CP1
PE1	6.80	954	134	1200	23.80	3.70	4CV	FRAN	CP1
PE3	5.80	1124	142	1200	21.40	3.70	5CV	FRAN	CP3
DA2	9.20	1360	170	1200	13.90	3.70	6CV	ETRA	CP3
PE9	8.70	1580	190	1200	11.20	3.70	6CV	FRAN	CP4
RE1	6.30	956	115	950	33.10	3.67	4CV	FRAN	CP1
RE3	6.30	1108	120	950	28.40	3.67	5CV	FRAN	CP2
RE4	5.80	1108	143	915	20.60	3.59	5CV	FRAN	CP2
FO9	7.90	1397	167	915	13.80	3.59	6CV	ETRA	CP3
RE8	8.70	1397	200	915	10.20	3.59	6CV	FRAN	CP4
SE4	8.80	1461	175	1200	14.70	3.63	6CV	ETRA	CP3
SE9	7.30	903	131	1088	23.40	3.46	4CV	ETRA	CP1
SZ2	6.40	993	145	400	18.40	3.58	4CV	ETRA	CP1
SZ3	6.50	1324	163	400	14.00	3.58	5CV	ETRA	CP3
TO1	6.10	999	150	202	19.50	3.70	4CV	ETRA	CP2
TO3	6.80	1295	170	202	15.00	3.70	5CV	ETRA	CP3
VW3	7.80	1272	170	1040	14.30	3.65	6CV	ETRA	CP3

TAB. 7 – *Données brutes relatives aux caractéristiques des voitures.*

NOM	4CV	5CV	6CV	FRAN	ETRA	CP1	CP2	CP3	CP4
AS2	5.386	-1.532	-2.853	-6.486	7.487	9.599	-4.236	-0.523	-3.839
CI4	5.386	-1.532	-2.853	-5.486	6.487	9.599	-4.236	-0.523	-3.839
PE6	5.386	-1.532	-2.853	-5.486	6.487	8.599	-3.236	-0.523	-3.839
FI3	5.386	-1.532	-2.853	-6.486	7.487	9.599	-4.236	-0.523	-3.839
FI5	5.386	-1.532	-2.853	-6.486	7.487	8.599	-3.236	-0.523	-3.839
FI8	4.386	-1.532	-1.853	-6.486	7.487	8.599	-4.236	-0.523	-2.839
FID	4.386	-1.532	-1.853	-6.486	7.487	8.599	-4.236	0.477	-3.839
FO1	5.386	-1.532	-2.853	-6.486	7.487	9.599	-4.236	-0.523	-3.839
RE7	4.386	-1.532	-1.853	-5.486	6.487	8.599	-4.236	-0.523	-2.839
NI1	5.386	-1.532	-2.853	-6.486	7.487	9.599	-4.236	-0.523	-3.839
OP1	5.386	-1.532	-2.853	-6.486	7.487	9.599	-4.236	-0.523	-3.839
PE1	5.386	-1.532	-2.853	-5.486	6.487	9.599	-4.236	-0.523	-3.839
PE3	4.386	-0.532	-2.853	-5.486	6.487	8.599	-4.236	0.477	-3.839
DA2	4.386	-1.532	-1.853	-6.486	7.487	8.599	-4.236	0.477	-3.839
PE9	4.386	-1.532	-1.853	-5.486	6.487	8.599	-4.236	-0.523	-2.839
RE1	5.386	-1.532	-2.853	-5.486	6.487	9.599	-4.236	-0.523	-3.839
RE3	4.386	-0.532	-2.853	-5.486	6.487	8.599	-3.236	-0.523	-3.839
RE4	4.386	-0.532	-2.853	-5.486	6.487	8.599	-3.236	-0.523	-3.839
FO9	4.386	-1.532	-1.853	-6.486	7.487	8.599	-4.236	0.477	-3.839
RE8	4.386	-1.532	-1.853	-5.486	6.487	8.599	-4.236	-0.523	-2.839
SE4	4.386	-1.532	-1.853	-6.486	7.487	8.599	-4.236	0.477	-3.839
SE9	5.386	-1.532	-2.853	-6.486	7.487	9.599	-4.236	-0.523	-3.839
SZ2	5.386	-1.532	-2.853	-6.486	7.487	9.599	-4.236	-0.523	-3.839
SZ3	4.386	-0.532	-2.853	-6.486	7.487	8.599	-4.236	0.477	-3.839
TO1	5.386	-1.532	-2.853	-6.486	7.487	8.599	-3.236	-0.523	-3.839
TO3	4.386	-0.532	-2.853	-6.486	7.487	8.599	-4.236	0.477	-3.839
VW3	4.386	-1.532	-1.853	-6.486	7.487	8.599	-4.236	0.477	-3.839

TAB. 8 – *Données qualitatives quantifiées.*

Summary

The processing of mixed data - quantitative and qualitative variables cannot be carry out directly by classical methods of data analysis. In this work, a factorial method which analyze simultaneously quantitative and qualitative data is described. The proposed Mixed Principal Component Analysis (MPCA) is a standardized principal component analysis of both quantitative variables and and quantified dummy variables associated to qualitative variables. For qualitative variables, the quantifications procedure is based on an orthogonal projection of configurations of statistical units in the individual-space, corresponding to Multivariate ANalysis of VAriance (MANOVA) coordinates. An example resulting from real mixed data illustrates the results of this method.