Université de Poitiers
Département de Mathématiques

Statistique descriptive, 1er semestre, année univ. 2009-2010 **Fiche 10**

AFC / ACM / Classification

Exercice 1

Un exemple d'étude en AFC

L'exemple concerne l'analyse d'un tableau de contingence qui croise 8 catégories socio-professionnelles (CSP) et 6 types de médias pour un échantillon de 11.130 "contacts médias" L'individu statistique sera pour nous le contact média.
Afin d'interpréter plus efficacement les représentations obtenues, on projettera en éléments supplémentaires certaines autres caractéristiques de la population telles que le sexe, l'âge, le niveau d'instruction.
Commenter et interpréter les résultats de l'analyse statistique présentés ci-après.

Les données sont les suivantes :

Profession	Radio	Television	Quot Nat	Quot Reg	Pr Mag	Pr TV
Agriculteur	96	118	2	71	50	17
Petit Patron	122	136	11	76	49	41
Prof. Cad. Sup.	193	184	74	63	103	79
Employe	511	593	57	217	172	306
Ouvier Qual.	385	457	42	174	104	220
Ouvrier Non Qual.	156	185	8	69	42	85
Inactif	1474	1931	181	852	642	782

Et pour les variables supplémentaires :

	Radio	Télévision	Quot Nat	Quot Reg	Pr Mag	Pr TV
Sexe						
Homme	1630	1900	285	854	621	776
Femme	1667	2069	152	815	683	938
Age						
15-24 ans	660	713	69	216	234	360
25-34 ans	640	719	84	230	212	380
35-39 ans	888	1000	130	429	345	466
50-64 ans	617	774	84	391	262	263
65 ans ou +	491	761	70	402	251	245
Instruction						
Primaire	908	1307	73	642	360	435
Secondaire	869	1008	107	408	336	494
Techn.Prof.	901	1035	80	140	311	504
Superieur	619	612	177	209	298	281

Les résultats relatifs à l'AFC sont les suivants.

Le tableau des fréquences :

	Radio	TV	Quot. Nat.	Quot. Reg.	Pr. Mag.	Pr. TV
Agr.	0.009	0.011	0.000	0.006	0.004	0.002
Petit Patron	0.011	0.012	0.001	0.007	0.004	0.004
Cad. Sup.	0.017	0.017	0.007	0.006	0.009	0.007
Empl.	0.046	0.053	0.005	0.019	0.015	0.027
Ouv. Qualif.	0.035	0.041	0.004	0.016	0.009	0.020
Ouv. Non Qualif	0.014	0.017	0.001	0.006	0.004	0.008
Inact.	0.132	0.173	0.016	0.077	0.058	0.070

Profils-ligne :

	Radio	TV	Quot. Nat.	Quot. Reg.	Pr. Mag.	Pr. TV
Agr.	0.271	0.333	0.006	0.201	0.141	0.048
Petit Patron	0.280	0.313	0.025	0.175	0.113	0.094
Cad. Sup.	0.277	0.264	0.106	0.091	0.148	0.114
Empl.	0.275	0.320	0.031	0.117	0.093	0.165
Ouv. Qualif.	0.279	0.331	0.030	0.126	0.075	0.159
Ouv. Non Qualif	0.286	0.339	0.015	0.127	0.077	0.156
Inact.	0.251	0.329	0.031	0.145	0.110	0.133

Profils-colonne :

	Radio	TV	Quot. Nat.	Quot. Reg.	Pr. Mag.	Pr. TV
Agr.	0.033	0.033	0.005	0.047	0.043	0.011
Petit Patron	0.042	0.038	0.029	0.050	0.042	0.027
Cad. Sup.	0.066	0.051	0.197	0.041	0.089	0.052
Empl.	0.174	0.165	0.152	0.143	0.148	0.200
Ouv. Qualif.	0.131	0.127	0.112	0.114	0.090	0.144
Ouv. Non Qualif	0.053	0.051	0.021	0.045	0.036	0.056
Inact.	0.502	0.536	0.483	0.560	0.552	0.511

Le tableau des valeurs propres est

	1	2	3	4	5
val. propres	0.01	0.01	0.00	0.00	0.00
part	0.62	0.33	0.03	0.01	0.00
part cumulée	0.62	0.95	0.99	1.00	1.00

Les coordonnées factorielles, contributions et qualités des individus actifs :

	Poids	QLT	INR	k=1	$\cos^2$	CTR	k=2	$\cos^2$	CTR
Agr.	32.00	961.00	169.00	137.00	150.00	41.00	318.00	811.00	409.00
Petit Patron	39.00	819.00	47.00	49.00	83.00	6.00	144.00	735.00	103.00
Cad. Sup.	63.00	998.00	548.00	-453.00	995.00	881.00	26.00	3.00	5.00
Empl.	167.00	932.00	74.00	12.00	15.00	2.00	-98.00	917.00	203.00
Ouv. Qualif.	124.00	883.00	72.00	39.00	113.00	13.00	-103.00	770.00	166.00
Ouv. Non Qualif	49.00	906.00	49.00	116.00	566.00	45.00	-90.00	340.00	50.00
Inact.	527.00	688.00	41.00	18.00	175.00	12.00	31.00	513.00	63.00

TAB. 1 – Les grandeurs ont été multipliées par 1000. Poids représente le poids de l'individu (PrL), QLT représente la qualité de représentation de l'individu dans le premier plan factoriel, INR la contribution de l'individu (PrL) à l'inertie. Puis axe factoriel par axe factoriel suivent les coordonnées, la qualité de la représentation et la contribution de l'indi à la construction de l'axe.

et pour les individus supplémentaires, les profils-ligne sont

	Radio	TV	Quot. Nat.	Quot. Reg.	Pr. Mag.	Pr. TV
Hommes	0.132	0.153	0.023	0.069	0.050	0.063
Femmes	0.135	0.167	0.012	0.066	0.055	0.076
15-24	0.053	0.058	0.006	0.017	0.019	0.029
25-34	0.052	0.058	0.007	0.019	0.017	0.031
35-39	0.072	0.081	0.010	0.035	0.028	0.038
50-64	0.050	0.062	0.007	0.032	0.021	0.021
>64	0.040	0.061	0.006	0.032	0.020	0.020
Inst. prim.	0.075	0.108	0.006	0.053	0.030	0.036
Inst. Second.	0.072	0.083	0.009	0.034	0.028	0.041
Techn. prof.	0.074	0.085	0.007	0.012	0.026	0.042
Inst. sup	0.051	0.051	0.015	0.017	0.025	0.023

La représentation des profils-ligne dans le premier plan factoriel est (quantités multipliée par 1000) :

	Poids	QLT	k=1	$\cos^2$	k=2	$\cos^2$
Hommes	490.00	2.00	-65.00	2.00	15.00	0.00
Femmes	510.00	1.00	41.00	1.00	-23.00	0.00
15-24	182.00	2.00	-17.00	0.00	-103.00	2.00
25-34	183.00	3.00	-32.00	0.00	-126.00	3.00
35-39	263.00	1.00	-41.00	0.00	-19.00	0.00
50-64	193.00	2.00	4.00	0.00	97.00	2.00
>64	179.00	4.00	42.00	0.00	134.00	3.00
Inst. prim.	307.00	6.00	110.00	4.00	83.00	2.00
Inst, Second.	266.00	1.00	-7.00	0.00	-45.00	1.00
Techn. prof.	245.00	9.00	-25.00	0.00	-187.00	9.00
Inst. sup	181.00	17.00	-306.00	17.00	-7.00	0.00

La représentation dans le premier plan factoriel des profils-colonne est (quantités multipliée par 1000) :

	Poids	QLT	INER	k=1	$\cos^2$	Contr.	k=2	$\cos^2$	Contr.
Radio	264.00	184.00	29.00	-7.00	17.00	1.00	-21.00	167.00	14.00
TV	324.00	904.00	35.00	48.00	902.00	51.00	-2.00	2.00	0.00
Quot. Nat.	34.00	996.00	482.00	-578.00	994.00	774.00	-28.00	2.00	3.00
Quot. Reg.	137.00	977.00	138.00	96.00	391.00	87.00	118.00	586.00	241.00
Pr. Mag.	104.00	938.00	140.00	-106.00	355.00	80.00	136.00	583.00	244.00
Pr. TV	137.00	963.00	177.00	26.00	23.00	7.00	-169.00	940.00	497.00

Enfin, la représentation dans le premier plan factoriel est :

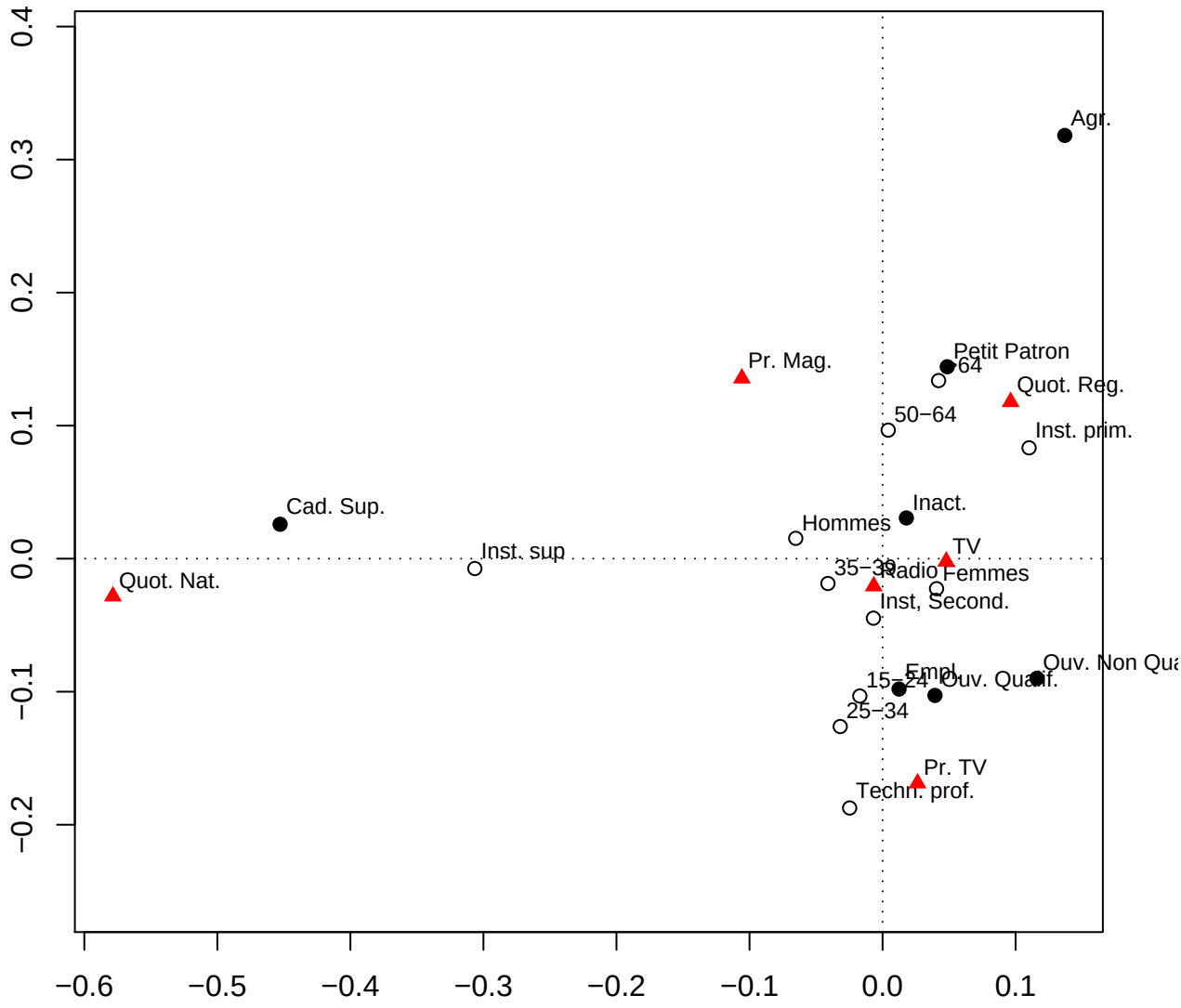

FIG. 1 – Représentation simultanée des deux nuages (PrL et PrC) dans le premier plan factoriel.

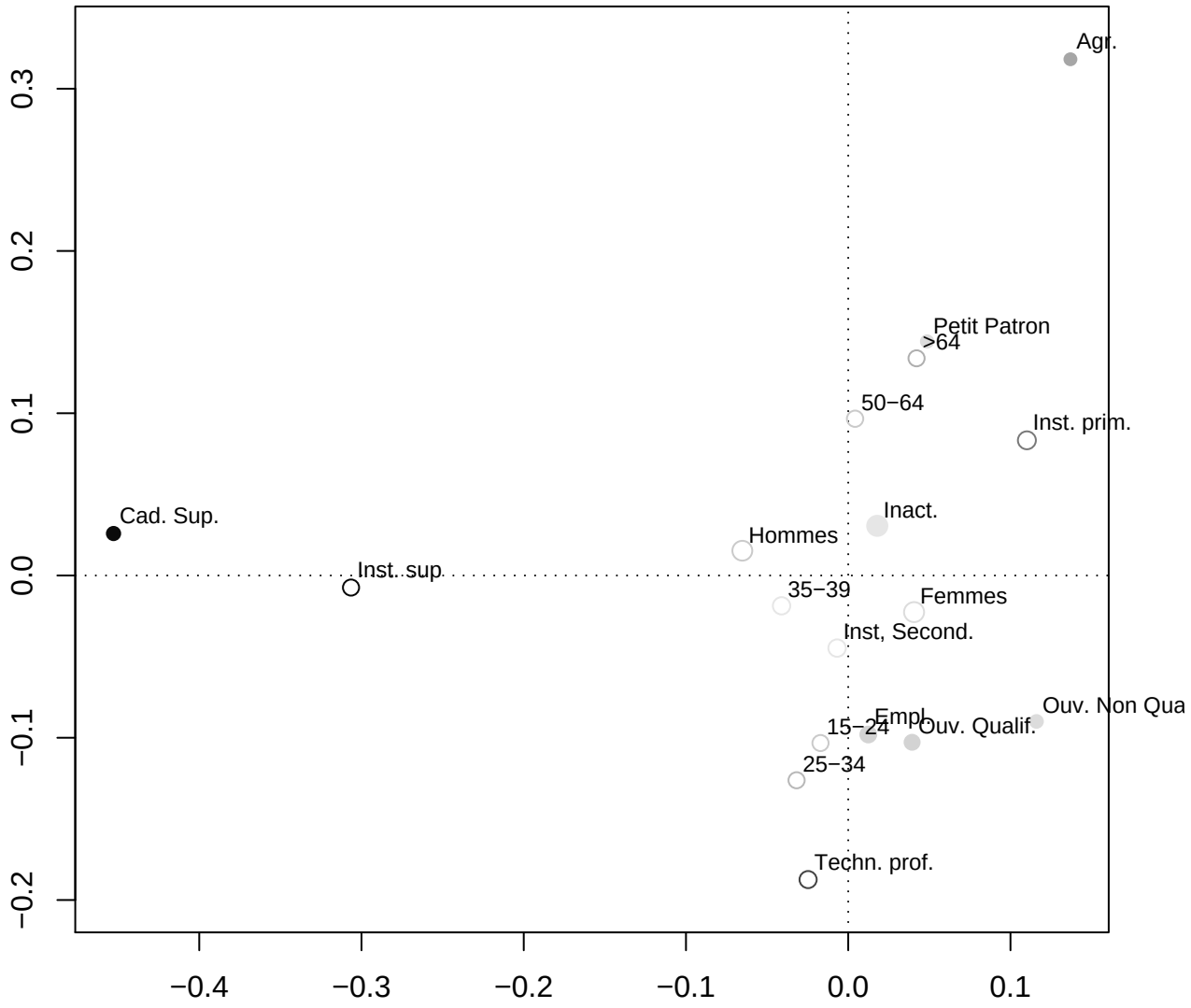

FIG. 2 – Représentation du nuage des PrL dans le premier plan factoriel.

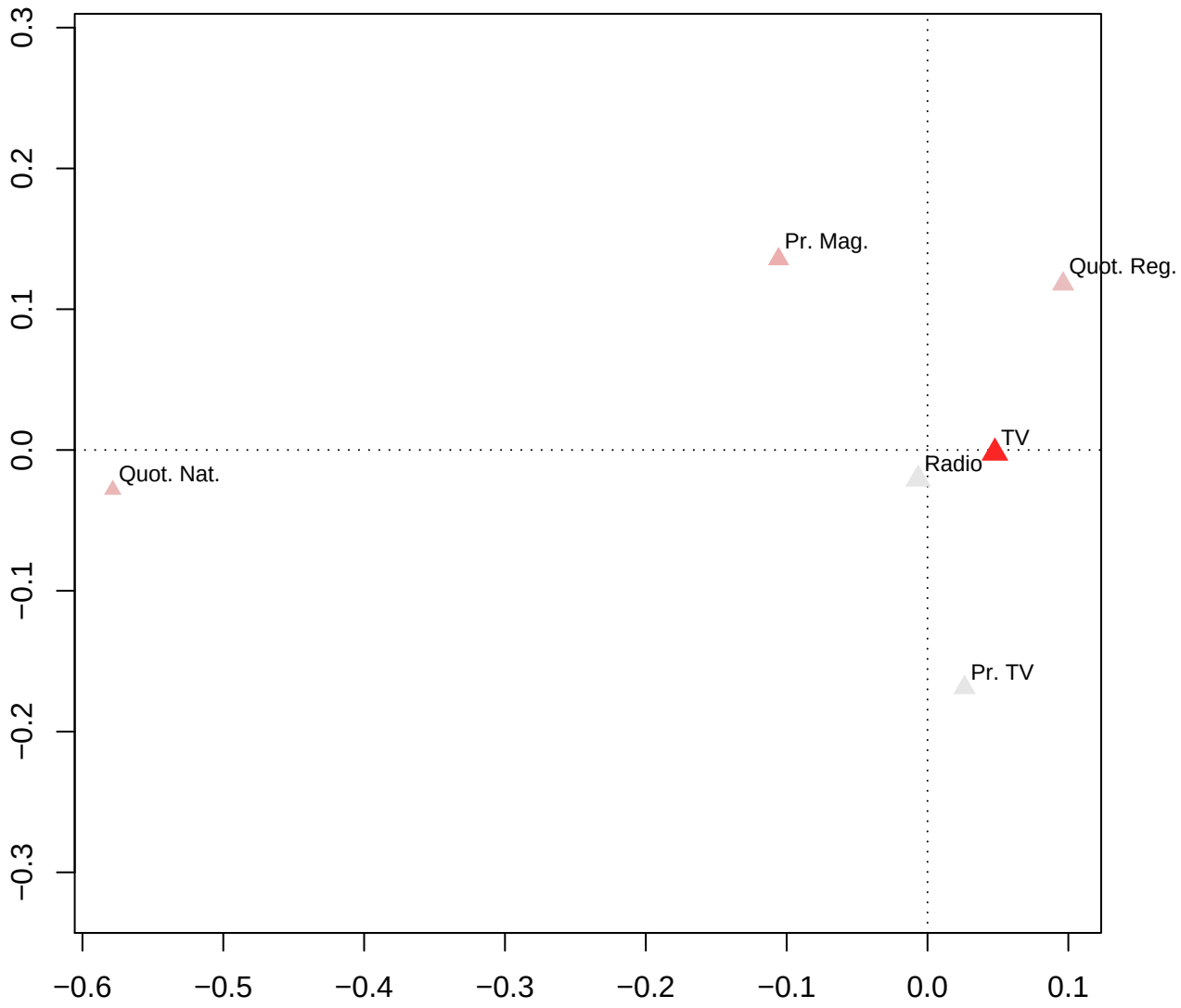

FIG. 3 – Représentation du nuage des PrC dans le premier plan factoriel.

Exercice 2

Un exemple d'étude en ACM

On considère les résultats d'une enquête réalisée en 1993 par *International Social Survey Programme* auprès de 871 individus à propos de leur attitude vis-à-vis de la science et de l'environnement. Quatre assertions sont proposées et pour chacune il y a 5 possibilités de réponse codées de 1 à 5, 1 signifie que l'on est totalement d'accord, 5 que l'on est totalement en désaccord, 3 que l'on est indécis. Par ailleurs, pour chaque individu est relevé son sexe (deux modalités), son âge (six modalités), son niveau d'études (six modalités).
Les résultats sont résumés ci-après :

A	B	C	D	sex	age	edu
1:119	1: 71	1:152	1: 60	1:427	1: 91	1: 38
2:322	2:174	2:316	2:232	2:444	2:210	2:378
3:204	3:205	3:197	3:202		3:158	3:242
4:178	4:281	4:154	4:226		4:146	4: 94
5: 48	5:140	5: 52	5:151		5:124	5: 49
					6:142	6: 70

L'analyse des correspondances multiples sur la matrice de Burt fournit les résultats suivants :

```
Principal inertias (eigenvalues):

      dim    value      %   cum%   scree plot
 [1,] 1      0.209196  18.6  18.6  *************************
 [2,] 2      0.185732  16.5  35.0  ***********************
 [3,] 3      0.103636   9.2  44.2  ************
 [4,] 4      0.093926   8.3  52.5  ***********
 [5,] 5      0.075997   6.7  59.3  ********
 [6,] 6      0.063468   5.6  64.9  ******
 [7,] 7      0.058835   5.2  70.1  ******
 [8,] 8      0.055202   4.9  75.0  *****
 [9,] 9      0.050836   4.5  79.5  *****
[10,] 10     0.048677   4.3  83.8  ****
[11,] 11     0.044032   3.9  87.7  ****
[12,] 12     0.038868   3.4  91.2  ***
[13,] 13     0.031642   2.8  94.0  **
[14,] 14     0.028599   2.5  96.5  **
[15,] 15     0.023354   2.1  98.6  *
[16,] 16     0.015687   1.4 100.0
[17,]        -------- -----
[18,] Total: 1.127684 100.0
```

Columns:

	name	mass	qlt	inr	k=1	cor	ctr	k=2	cor	ctr
1	A.1	34	445	55	840	391	53	314	54	8
2	A.2	92	169	38	250	136	13	-123	33	3
3	A.3	59	344	47	-204	47	5	-517	298	36
4	A.4	51	350	50	-533	258	32	318	92	12
5	A.5	14	401	60	-913	170	25	1064	231	36
6	B.1	20	621	62	1338	519	80	590	101	16
7	B.2	50	158	47	293	80	9	-287	77	10
8	B.3	59	227	45	158	29	3	-415	198	24
9	B.4	81	210	41	-327	185	19	-121	25	3
10	B.5	40	722	60	-619	229	34	908	493	77
11	C.1	44	732	60	987	632	93	392	100	16
12	C.2	91	164	38	113	27	3	-255	137	14
13	C.3	57	296	48	-283	84	10	-450	212	27
14	C.4	44	345	52	-617	289	37	274	57	8
15	C.5	15	471	60	-671	99	15	1300	372	59
16	D.1	17	251	56	551	83	11	785	168	25
17	D.2	67	14	42	-101	14	1	-3	0	0
18	D.3	58	303	48	-176	33	4	-499	269	34
19	D.4	65	25	43	-101	14	1	-91	11	1
20	D.5	43	272	50	324	81	10	496	191	25

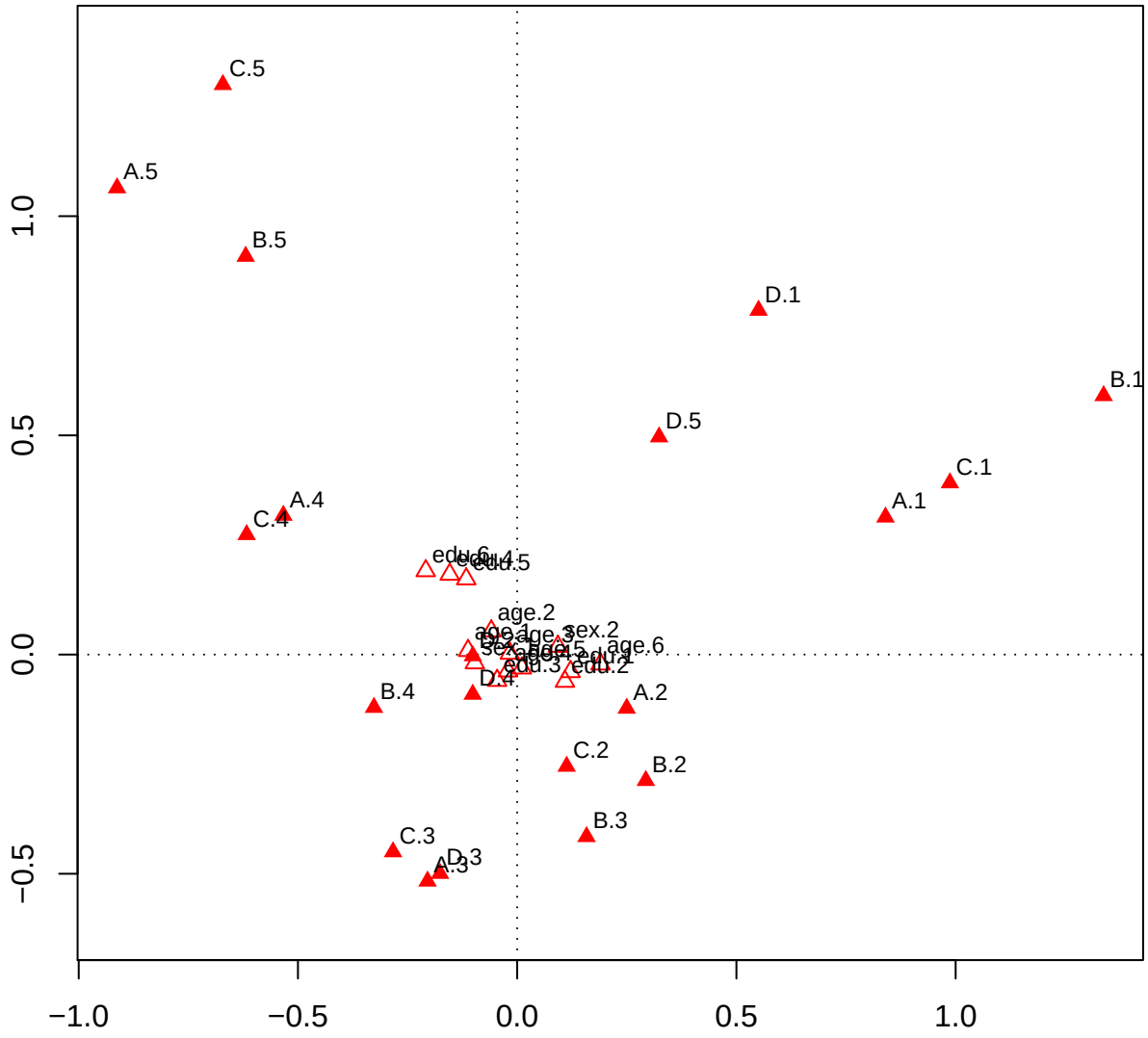

FIG. 4 – Représentation du nuage des colonnes dans le premier plan factoriel. Les variables supplémentaires sont représentées.

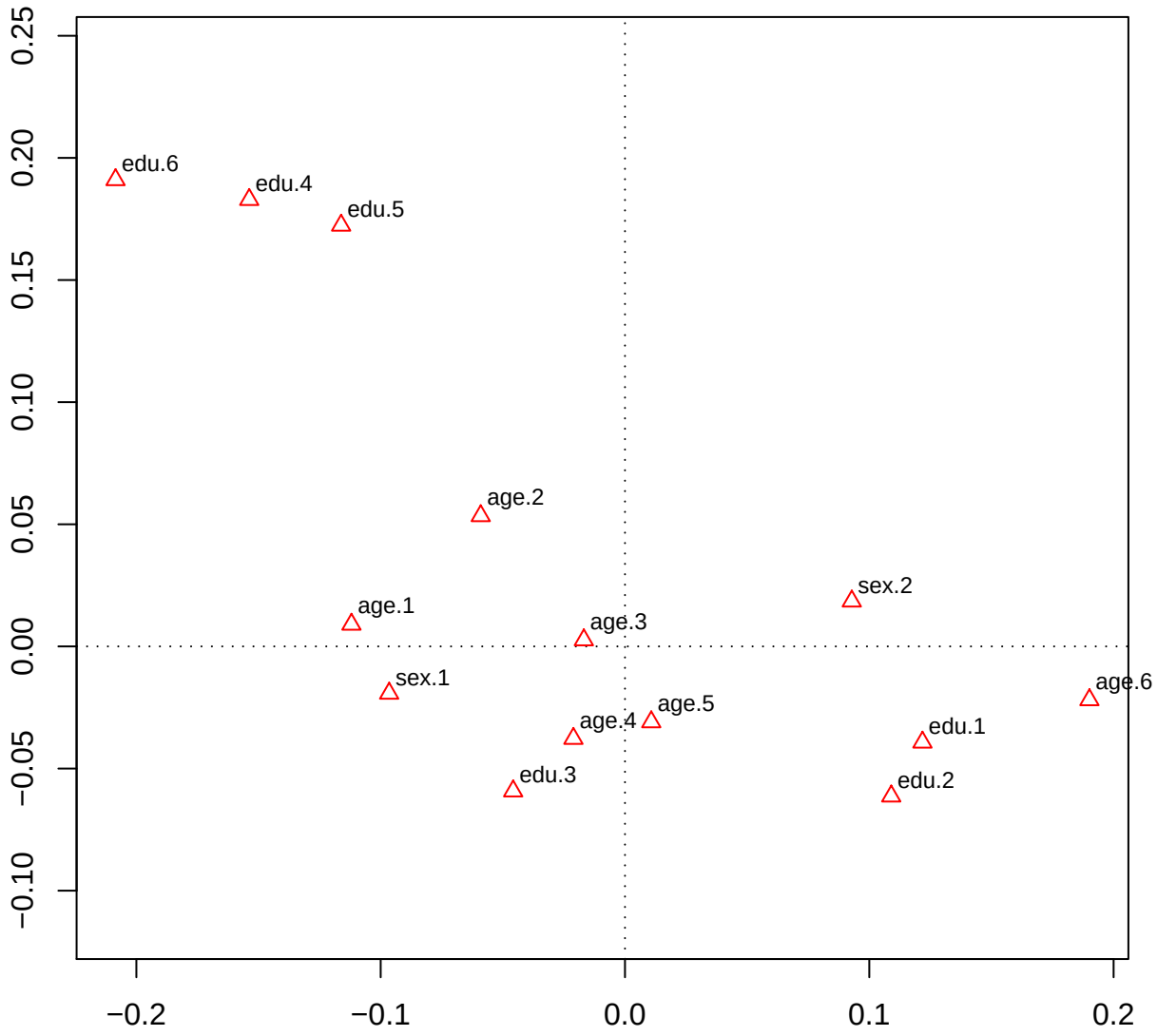

FIG. 5 – Représentation dans le premier plan factoriel uniquement des variables supplémentaires.

Exercice 3

Algorithme d'agrégation autour des centres mobiles

Soit un nuage $\mathcal{N}$ de n points $x_1, \ldots, x_n$ dans $\mathbb{R}^n$ muni de la distance euclidienne. On souhaite résumer ce nuage de points par k points $z_1, \ldots, z_k$ avec $k \leq n$ en associant chacun des points du nuage $\mathcal{N}$ à un unique point z_i. L'ensemble des points x_k associés à z_i est alors appelé la i-ème classe et est noté C_i, et on appelle z_i le représentant de la classe C_i. Le critère retenu est de minimiser l'erreur quadratique moyenne

$$\sum_{i=1}^{k} \sum_{x \in C_i} d(x, z_i)^2$$

obtenue lorsque l'on remplace chaque individu par le représentant de sa classe.

1) Soit $\{(z_i, C_i), 1 \leq i \leq k\}$ une solution proposée. On pose z'_i le centre de gravité de la i-ème classe. Montrer que $\{z'_i, C_i, 1 \leq i \leq k\}$ est une meilleure solution. Cela montre que le meilleur représentant d'une classe est son centre de gravité.

2) Soit $\{(z_i, C_i), 1 \leq i \leq k\}$ une solution proposée. Soit $\{C'_i, 1 \leq i \leq k\}$ la partition obtenue en réaffectant chaque x_k au point z_i dont il est le plus proche. Montrer que $\{(z_i, C'_i), 1 \leq i \leq k\}$ est une meilleure solution.

3) Montrer que l'*algorithme d'agrégation autour des centres mobiles* s'arrête au bout d'un temps fini. Quelles propriétés a la solution proposée ?