TP2 : enquête sur les clients d'une banque (Correction)

On veut analyser des données relatives à une enquête réalisée sur 810 clients d'une banque. On s'intéresse tout d'abord aux variables suivantes

- `solde` : solde du compte courant (`p4` (>20 k€), `p3` (12–20 k€), `p2` (4–12 k€), `p1` (0–4k€), `n1` (-4–0 k€), `n2` (< - 4 k€)).
- `interdit` : interdit de chéquier (`oui` ou `non`)
- `eparlog` : plan d'épargne logement (`nul` (non), `fai` (<20 k€), `for` (>20 k€))
- `eparliv` : livret d'épargne (`nul` (non), `fai` (<20 k€), `for` (>20 k€))
- `credcon` : crédits à la consommation (`nul` (non), `fai` (<20 k€), `for` (>20 k€))
- `portit` : portefeuille de titres (`nul` (non), `fai` (<20 k€), `moy` (20–100 k€), `for` (>100 k€))

Le code pour charger les données est le suivant :

```
> require(ade4)
> source("fonctions.R")
> data(banque)
> banque1=subset(banque,select=c(soldevu,interdit, eparlog, eparliv, credcon, porttit))
> # renomme soldevu en solde
> colnames(banque1)[1]="solde"
```

1 Les données

Question 1 *Calculez le tableau de Burt des données ci-dessus avec la fonction* `acm.burt`.

On utilise directement la fonction `acm.burt` sans difficulté particulière (même s'il est trop large pour s'afficher ici).

```
> acm.burt(banque1, banque1)
```

	solde.p4	solde.p3	solde.p2	solde.p1	solde.n1	solde.n2	interdit.non	interdit.oui	eparlog.for	eparlog.fai	eparlog.nul	eparliv.f
solde.p4	95	0	0	0	0	0	92	3	19	2	74	
solde.p3	0	69	0	0	0	0	68	1	12	11	46	
solde.p2	0	0	145	0	0	0	140	5	21	13	111	
solde.p1	0	0	0	271	0	0	257	14	9	9	253	
solde.n1	0	0	0	0	162	0	132	30	1	9	152	
solde.n2	0	0	0	0	0	68	63	5	2	0	66	
interdit.non	92	68	140	257	132	63	752	0	62	44	646	
interdit.oui	3	1	5	14	30	5	0	58	2	0	56	
eparlog.for	19	12	21	9	1	2	62	2	64	0	0	
eparlog.fai	2	11	13	9	9	0	44	0	0	44	0	
eparlog.nul	74	46	111	253	152	66	646	56	0	0	702	
eparliv.for	17	3	10	8	5	1	42	2	12	3	29	
eparliv.fai	24	20	40	31	18	11	142	2	32	15	97	
eparliv.nul	54	46	95	232	139	56	568	54	20	26	576	
credcon.nul	87	56	116	251	135	40	633	52	54	37	594	
credcon.fai	2	8	18	9	20	11	63	5	5	1	62	
credcon.for	6	5	11	11	7	17	56	1	5	6	46	
porttit.nul	48	50	109	219	146	58	583	47	36	29	565	
porttit.fai	7	7	17	27	7	4	66	3	8	8	53	
porttit.moy	14	8	14	17	5	3	58	3	11	5	45	
porttit.for	26	4	5	8	4	3	45	5	9	2	39	

Question 2 *Calculer avec la fonction* `chisq.test` *le* χ^2 *des variables* `eparlog` *et* `credcon`. *Que pouvez-vous en dire ? Même question avec les variables* `eparlog` *et* `eparliv`.

On calcule directement

```
> chisq.test(banque1$eparlog, banque1$credcon)

        Pearson's Chi-squared test

data:  banque1$eparlog and banque1$credcon
X-squared = 5.1634, df = 4, p-value = 0.2709
```

On peut vérifier le nombre de degrés de liberté de la table de contingence : comme les deux variables ont 3 modalités, on trouve $(3-1)\times(3-1)=4$ degrés de liberté. La p-value est bien trop grande pour rejeter l'hypothèse d'indépendance. En conclusion, il n'est pas possible d'établir un lien entre la possession d'un plan épargne logement et celle d'un crédit à la consommation, même à un seuil de 5% d'erreur.

Par contre, si on compare `eparlog` et `eparliv`, on obtient

```
> chisq.test(banque1$eparlog, banque1$eparliv)

        Pearson's Chi-squared test

data:  banque1$eparlog and banque1$eparliv
X-squared = 95.007, df = 4, p-value < 2.2e-16
```

Dans ce cas, l'hypothèse d'indépendance est très clairement rejetée et on peut considérer qu'il y a un lien entre la possession d'un plan d'épargne logement et celle d'un livret d'épargne.

2 Analyse des correspondances multiples

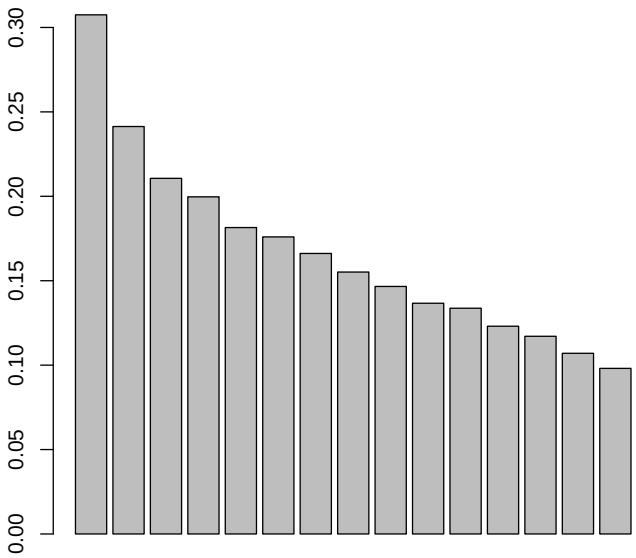
Question 3 *Réaliser une ACM sur les données ci-dessus à l'aide de la fonction* ***dudi.acm****.*

```
> acm1 = dudi.acm(banque1,scannf=F,nf=3)
> # cette commande ci-dessous est juste pour s'assurer des signes.
> acm1 = dudi.fixsigns(acm1, sign.co=c(1,-1,1))
> # des choses qui peuvent être utiles dans l'analyse.
> inert1 = inertia.dudi(acm1,c=T,r=T)
```

Question 4 *Combien d'axes propres faut-il conserver d'après les règles habituelles ? Quelle est la proportion de l'inertie expliquée par le sous-espace propre correspondant (on pourra utiliser la fonction* ***inertia.dudi****) ?*

L'histogramme des valeurs propres est

```
> barplot(acm1$eig)
```



La règle la plus classique consiste à conserver les axes associés aux valeurs propres supérieures à $1/p$, où p est le nombre de variables actives (6 ici). Ici la décroissance des valeurs propres est très faible ce qui va nous conduire à une analyse médiocre. On devrait conserver les 5 axes qui nous sont donnés, et plus encore, ce qui n'est pas très satisfaisant. On peut par contre remarquer que les deux premières valeurs propres sont un peu à l'écart des suivantes.

On choisit donc ici de se contenter des deux premiers axes. On est typiquement ici dans un cas où il y a très peu de dépendances entre les variables.

Pour calculer l'inertie expliquée, on peut utiliser `inertia.dudi` (qu'on invoque ici comme `inertia` parce que l'objet acm1 est de classe `"dudi"` :

```
> inertia(acm1)

Inertia information:
Call: inertia.dudi(x = acm1)

Decomposition of total inertia:
     inertia     cum  cum(%)
Ax1  0.30748  0.3075   12.30
Ax2  0.24133  0.5488   21.95
Ax3  0.21062  0.7594   30.38
```

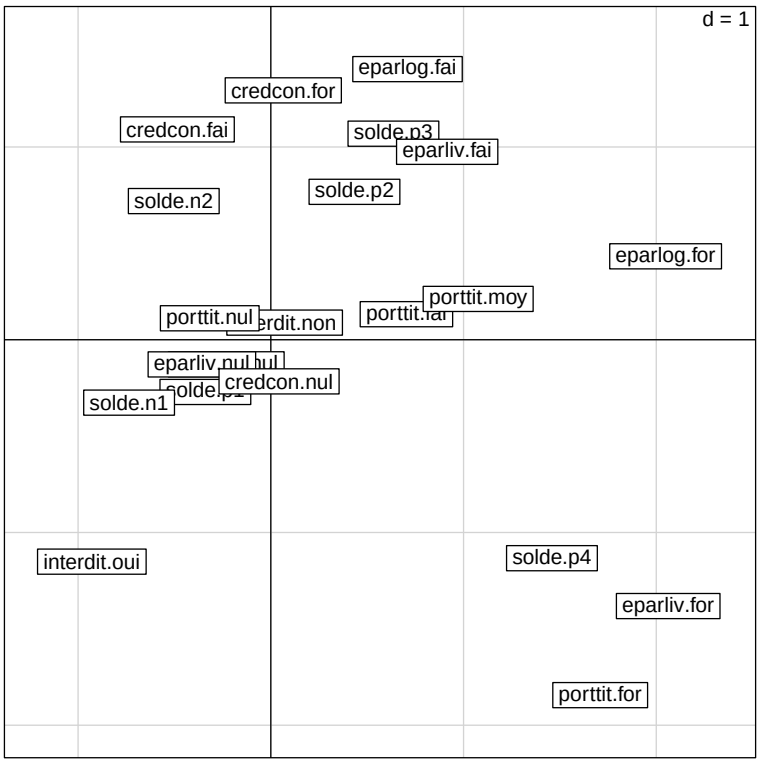
```
Ax4  0.19969  0.9591   38.36
Ax5  0.18148  1.1406   45.62
Ax6  0.17594  1.3165   52.66
Ax7  0.16614  1.4827   59.31
Ax8  0.15513  1.6378   65.51
Ax9  0.14660  1.7844   71.38
Ax10 0.13666  1.9211   76.84
Ax11 0.13370  2.0548   82.19
Ax12 0.12306  2.1778   87.11
Ax13 0.11709  2.2949   91.80
Ax14 0.10701  2.4019   96.08
Ax15 0.09807  2.5000  100.00
```

Avec deux axes, on a donc un peu plus de 20% de l'inertie. Ces résultats ne sont pas bons, mais on n'y peut pas grand-chose.

Question 5 *Faites une représentation des catégories. Quelles sont celles qui déterminent les deux premiers axes principaux ? (on détaillera les critères et on cherchera à être précis dans la réponse).*

La représentation des catégories est obtenue par

```
> s.label(acm1$co)
```



On va utiliser ici la méthode sur les coordonnées, même si on pourrait travailler sur les contributions et les poids. La contribution de chaque catégorie (de coordonnée a_j) à un axe factoriel (associé à la valeur propre μ) s'écrit

$$\frac{n_j}{np}\frac{(a_j)^2}{\mu}$$

et doit être comparée à son poids n_j/np. Si choisit donc un facteur 3 pour éviter d'avoir trop de variables, un calcul simple montre qu'il faut s'intéresser aux modalités dont les coordonnées sur l'axe vérifient

$$|a_j| > \sqrt{3\mu}.$$

On peut obtenir les limites par

```
> round(sqrt(3*acm1$eig[1:2]),2)
```

```
[1] 0.96 0.85
```

Les résultats obtenus sont les mêmes que si on avait raisonné directement sur les contributions. On classe les éléments par coordonnée décroissante.

Axe 1

$\ominus$	$\oplus$
[`interdit.oui` (-0.93)]	`eparliv.for` (2.06)
	`eparlog.for` (2.05)
	`porttit.for` (1.71)
	`solde.p4` (1.46)
	`porttit.moy` (1.08)
	[`eparliv.fai` (0.92)]

Axe 2

$\ominus$	$\oplus$
`porttit.for` (-1.84)	`eparlog.fai` (1.40)
`eparliv.for` (-1.38)	`credcon.for` $(1..29)$
`interdit.oui` (-1.15)	`credcon.fai` (1.09)
`solde.p4` (-1.13)	`solde.p3` (1.07)
	`eparliv.fai` (0.98)

Question 6 *Interprétez les axes et identifiez les difficultés.*

Le premier axe correspond à droite aux clients qui ont des moyens importants : épargne logement, livret, portefeuille de titres et solde élevé sur le compte courant. De manière peu étonnante, on trouve à l'opposé les interdits de chéquier.

Le second axe est plus difficile à interpréter : d'un côté on trouve certes des gens qui ont des moyens raisonnables (un peu d'épargne et d'argent sur leur compte, mais des crédits à la consommation) ; mais de l'autre, on mélange des clients avec de gros moyens et des interdits de chéquier. Il y a là une difficulté d'interprétation.

2.1 Catégories supplémentaires

On cherche à préciser les caractéristiques des axes en termes de type de client. On s'intéresse donc aux variables supplémentaires suivantes :

- `age` : âge du client (`ai25` [18, 25[, `ai35` [25, 35[, `ai45` [35, 45[, `ai55` [45, 55[, `ai75` [55, 75[) ;
- `sexe` : sexe du client (`hom` ou `fem`) ;
- `csp` : catégorie socio-professionnelle : agriculteur (`agric`), artisan (`artis`), cadre supérieur (`cadsu`), profession intermédiaire (`inter`), employé (`emplo`), ouvrier (`ouvri`), retraité (`retra`) , inactif (`inact`) et étudiant (`etudi`).

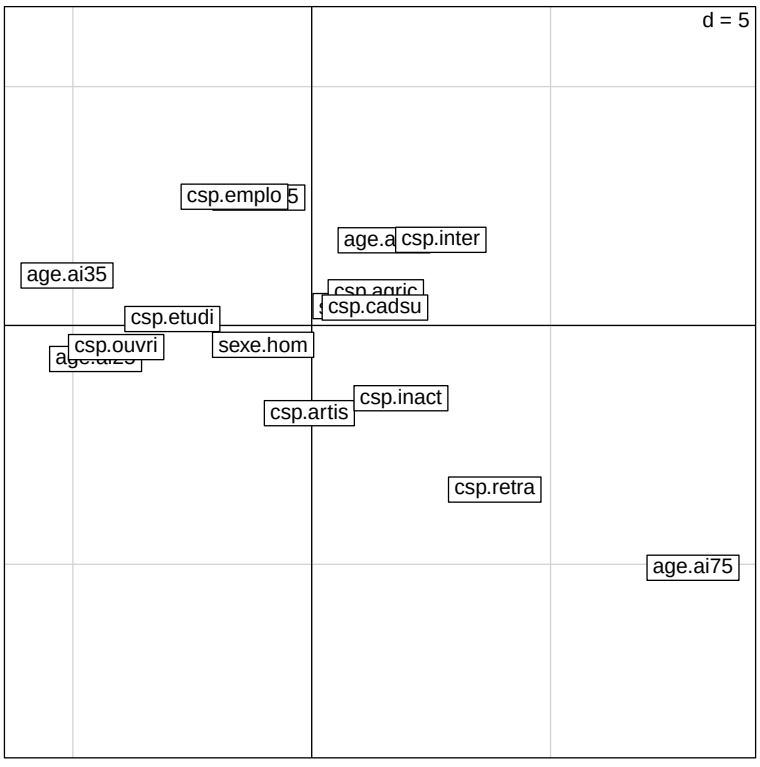
Question 7 *Utilisez `acm.suppl` du fichier `fonctions.R` pour calculer les effectifs et les valeurs tests correspondant aux individus supplémentaires.*

Le calcul se fait en deux étapes : d'abord, créer la table des variables supplémentaires en utilisant encore `subset` ; puis appliquer `acm.suppl`.

```
> banque1.suppl=subset(banque,select=c(age,sexe,csp))
> suppl1 = acm.suppl(acm1,banque1.suppl)
```

Question 8 *Vérifier si on peut utiliser les valeurs-test. Complétez l'interprétation des deux premiers axes grâce à ces valeurs.*

```
> s.label(suppl1$test)
```



Les valeurs test permettent de savoir si des catégories supplémentaires sont corrélées de manière significative avec les axes principaux. On peut les utiliser si

- on les utilise sur des variables qui n'ont pas pris part à l'analyse : c'est le cas ici ;
- les effectifs des catégories sont assez importants : ici elles sont $\geqslant 29$. Il n'y a guère que `agric` qui pourrait poser un problème. Les effectifs des catégories sont en effet

```
> summary(banque1.suppl, maxsum=20)
```

```
   age       sexe        csp
 ai25: 90   hom:558   agric: 29
 ai35:156   fem:252   artis: 48
 ai45:212             cadsu:103
 ai55:174             inter:102
 ai75:178             emplo:151
                      ouvri:183
                      retra: 52
                      inact: 85
                      etudi: 57
```

On considérera une valeur comme significative si elle est supérieure à 2 ou 3 en valeur absolue.

Axe 1

$\ominus$	$\oplus$
`age.ai35` (-5.12)	`age.ai75` (7.98)
`age.ai25` (-4.52)	`csp.retra` (3.83)
`csp.ouvri` (-4.09)	[`csp.inter` (2.70)]
[`csp.etudi` (2.92)]	

Axe 2

$\ominus$	$\oplus$
`age.ai75` (-5.07)	[`csp.empl` (2.70)]
`csp.retra` (-3.44)	[`ai.45` (2.68)]

*

L'axe 1 oppose les retraités âgés, et éventuellement les professions intermédiaires (mais pourquoi ?) qui ont des moyens financiers plus importants, d'une part, et des moins de 35 ans, en particulier ouvriers ou peut-être étudiants, qui peuvent avoir des problèmes de fin de mois. Le second axe, lui, oppose des employés de 35-45 ans avec des moyens « moyens » aux mêmes retraités âgés.

Une manière de mieux interpréter le graphique est peut-être de regarder les deux axes en même temps et de considérer que les retraités âgés avec un compte en banque garni, un bon portefeuille de titres et un bon livret d'épargne sont en bas à gauche. On remarquera qu'ils n'ont pas d'épargne logement forte, puisque ce problème est probablement résolu à leur âge. Il y a ici un effet Guttman, où les ages sont répartis selon un parabole retournée.

Enfin, on remarquera qu'il n'y a pas vraiment de différence entre les femmes et les hommes.