TD2 : Budget-temps (suite)

(Correction)

```
> require(ade4)
> source("fonctions.R")
> budget<-read.table("budget.dat")
> budget1<-budget[,1:10]
> pca1<-dudi.pca(budget1, scannf=F, nf=5)
> pca1 = dudi.fixsigns(pca1, sign.li=c(-1,-1,-1,1,-1))
```

On poursuit l'étude des données du TD1. En particulier, les questions 2 et 4 remplacent les questions 3 et 4 du TD précédent grâce aux nouveaux outils vus en cours.

On rappelle que les 10 variables numériques sont le temps passé en : PROFession, TRANsport, MENAge, ENFAnts, COURses, TOILette, REPAs, SOMMeil, TÉLÉ et LOISirs. Les temps sont en centièmes d'heure et le total d'une ligne (sur ces 10 variables numériques) est 2400 (24 heures).

Les 4 variables catégorisées sont :

— le SEXe (1=Hommes, 2=Femmes),
— l'ACTivité (1=Actifs, 2=Non Act., 9=Non précisé),
— l'état CIVil (1=Célibataires, 2=Mariés, 9=Non précisé),
— le PAYs (1=USA, 2=Pays de l'Ouest, 3=Pays de l'Est, 4=Yougoslavie).

Le code suivant est utilisé pour identifier les lignes : H : Hommes, F : Femmes, A : Actifs, N : Non Actifs, M : Mariés, C : Célibataires, U : USA, W : Pays de l'Ouest sauf USA, E : Est sauf Yougoslavie, Y : Yougoslavie.

On fait une ACP sur variables centrées-réduites des 10 variables numériques. On donne ci-dessous pour les 5 premiers axes les coordonnées des individus et leur contribution aux composantes principales (en %) ainsi que la corrélation des variables avec les composantes principales. Enfin on fournit les valeurs propres associées aux axes et la projection des individus sur le premier plan principal (gauche) et sur le plan (3,4) (droite).

Coordonnées des individus

```
> round(pca1$li, 2)
```

	Axis1	Axis2	Axis3	Axis4	Axis5
HAU	-1.77	-0.69	-1.87	0.58	-0.85
FAU	-0.17	-2.22	-0.66	0.44	1.25
FNU	4.05	-2.28	-1.06	-0.52	-1.04
HMU	-1.78	-0.29	-1.89	0.73	-1.04
FMU	2.61	-2.29	-0.80	0.11	-0.37
HCU	-1.50	-1.89	-1.36	-0.78	-0.35
FCU	-0.47	-2.84	-1.30	-0.15	1.62
HAW	-1.18	2.37	-1.12	-0.05	0.23
FAW	0.31	1.50	-0.27	0.94	1.25
FNW	4.32	1.63	-0.89	-0.14	-0.23
HMW	-1.13	2.46	-1.29	0.15	0.22
FMW	3.13	1.99	-0.59	0.73	-0.35
HCW	-1.37	2.57	-0.53	-1.02	-0.29
FCW	1.10	1.66	-0.54	-1.50	1.43
HAY	-2.16	0.24	0.71	-0.24	-0.32
FAY	-1.00	-0.18	1.62	2.13	-0.04
FNY	3.54	0.38	1.64	-0.53	-0.28
HMY	-2.22	0.21	0.48	-0.11	-0.40
FMY	1.54	0.22	1.62	1.17	-0.44
HCY	-2.14	-0.58	1.61	-2.18	-0.55
FCY	-0.34	-0.42	1.49	-1.02	0.12
HAE	-2.15	0.07	0.13	0.32	-0.15
FAE	-0.99	-0.59	1.37	2.04	0.27
FNE	3.92	-0.05	0.67	-0.98	0.00
HME	-2.08	0.17	-0.43	0.79	-0.22
FME	0.49	-0.20	1.18	2.01	0.04
HCE	-2.53	-0.15	1.06	-1.96	-0.35
FCE	-0.06	-0.80	1.00	-0.97	0.84

Contributions des individus

```
> inert1=inertia.dudi(pca1,r=T)
> colnames(inert1$row.abs)=paste0("Axis",1:5)
> round(inert1$row.abs,1)
```

	Axis1	Axis2	Axis3	Axis4	Axis5
HAU	2.4	0.8	9.5	1.0	5.6
FAU	0.0	8.3	1.2	0.6	11.9
FNU	12.8	8.7	3.0	0.8	8.2
HMU	2.5	0.1	9.6	1.6	8.2
FMU	5.3	8.8	1.7	0.0	1.0
HCU	1.8	6.0	5.0	1.8	1.0
FCU	0.2	13.6	4.5	0.1	19.9
HAW	1.1	9.4	3.4	0.0	0.4
FAW	0.1	3.8	0.2	2.7	11.9
FNW	14.5	4.5	2.1	0.1	0.4
HMW	1.0	10.2	4.5	0.1	0.4
FMW	7.6	6.7	0.9	1.6	0.9
HCW	1.5	11.1	0.7	3.1	0.6
FCW	0.9	4.6	0.8	6.7	15.6
HAY	3.6	0.1	1.4	0.2	0.8
FAY	0.8	0.1	7.1	13.6	0.0
FNY	9.7	0.2	7.2	0.8	0.6
HMY	3.8	0.1	0.6	0.0	1.2
FMY	1.8	0.1	7.1	4.1	1.5
HCY	3.5	0.6	7.0	14.2	2.3
FCY	0.1	0.3	6.0	3.1	0.1
HAE	3.6	0.0	0.0	0.3	0.2
FAE	0.8	0.6	5.0	12.4	0.5
FNE	12.0	0.0	1.2	2.8	0.0
HME	3.4	0.0	0.5	1.9	0.4
FME	0.2	0.1	3.8	12.1	0.0
HCE	5.0	0.0	3.1	11.5	0.9
FCE	0.0	1.1	2.7	2.8	5.4

Corrélations variables/composantes

```
> round(pca1$co,2)
```

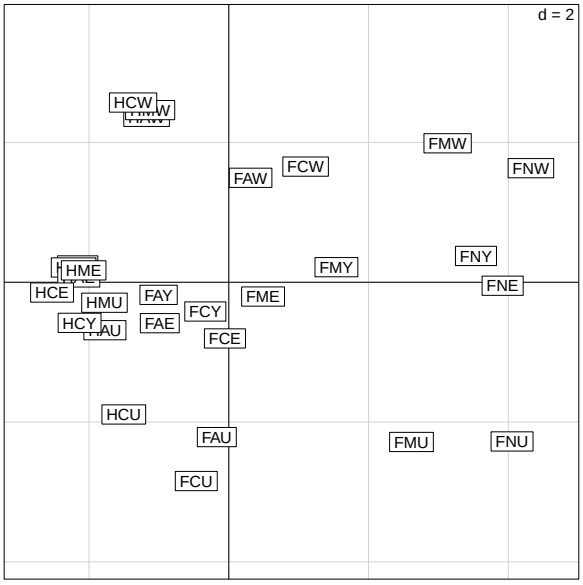
	Comp1	Comp2	Comp3	Comp4	Comp5
PROF	-0.98	0.12	-0.08	0.07	0.10
TRAN	-0.98	-0.06	-0.01	0.05	-0.11
MENA	0.90	-0.02	0.36	0.21	0.00
ENFA	0.87	-0.18	0.08	0.29	-0.19
COUR	0.56	-0.76	0.00	-0.12	-0.09
TOIL	0.08	-0.82	-0.30	-0.06	0.45
REPA	0.59	0.67	-0.43	0.01	0.00
SOMM	0.64	0.57	-0.19	-0.31	0.31
TELE	0.10	-0.19	-0.93	0.15	-0.24
LOIS	0.09	-0.11	0.03	-0.96	-0.22

Valeurs propres

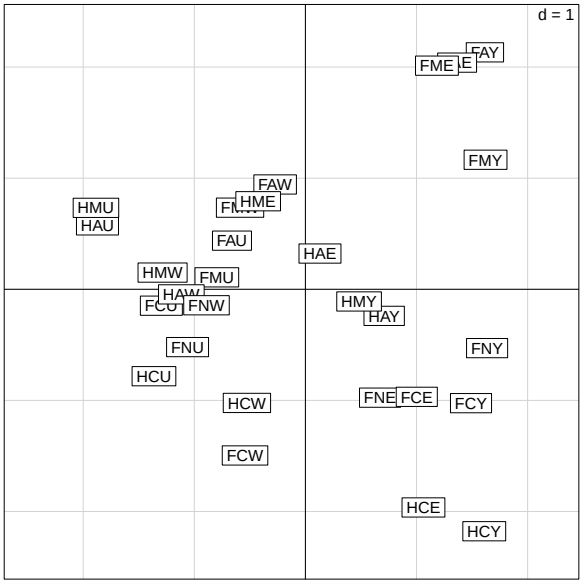
```
> eig = pca1$eig
> eig[2] = NA
> matrix(round(eig,4),length(eig),1,dimnames=
```

```
Axis1 4.5887
Axis2     NA
Axis3 1.3210
Axis4 1.1953
Axis5 0.4684
Axis6 0.1990
Axis7 0.0468
Axis8 0.0371
Axis9 0.0239
```

```
> s.label(pca1$li)
```




```
> s.label(pca1$li,3,4)
```



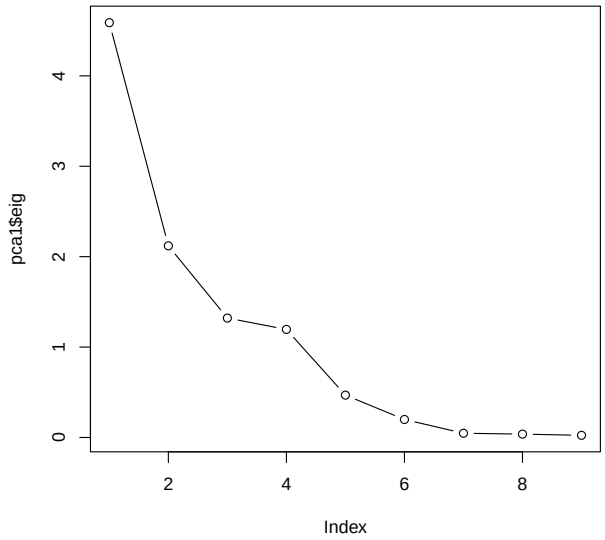
Question 1 *Calculer la valeur propre manquante (`NA`). Tracer la courbe des valeurs propres. Combien faut-il en retenir ?*

La première chose à remarquer est que l'on a que 9 valeurs propres, y compris celle qui manque. Pourquoi ? Parce que la somme des variables d'origine fait exactement 24h ; les variables ne sont donc pas indépendantes (on peut exprimer une des variables comme 2400 moins la somme des autres) et le rang de la matrice $\mathbf{R}$ est donc 9 plutôt que 10. En fait, la 10$^{\text{e}}$ valeur propre existe, elle est nulle.

On sait que la somme des valeurs propres est 10. Comme la somme des valeurs données ci-dessus est 7.8802, la seconde valeur propre vaut 2.1198.

L'histogramme des valeurs propres est représenté ci-dessous :

```
> plot(pca1$eig,type="b")
```



La règle de Kaiser conduit à retenir les valeurs propres qui sont supérieures à 1, ce qui signifie ici les 4 premières.

Question 2 *Pour chacune des 4 premières composantes principales, donner la liste des individus qui contribuent à l'axe de manière significative.*

On regarde les contributions aux axes, dont on veut qu'elles soient supérieures à 2 fois le poids. Comme il y a 28 individus, on veut des contributions supérieures à 7.14% . On fait bien attention de séparer les coordonnées positives des coordonnées négatives en se reportant à la projection des individus.

Axe 1

⊖	⊕
	FNW (14.5)
	FNU (12.8)
	FNE (12.0)
	FNY (9.7)
	FMW (7.6)

Axe 2

⊖	⊕
FCU (13.6)	HCW (11.1)
FMU (8.8)	HMW (10.2)
FNU (8.7)	HAW (9.4)
FAU (8.3)	

Axe 3

⊖	⊕
HMU (9.6)	FNY (7.2)
HAU (9.5)	[FMY (7.1)]
	[FAY (7.1)]
	[HCY (7.0)]

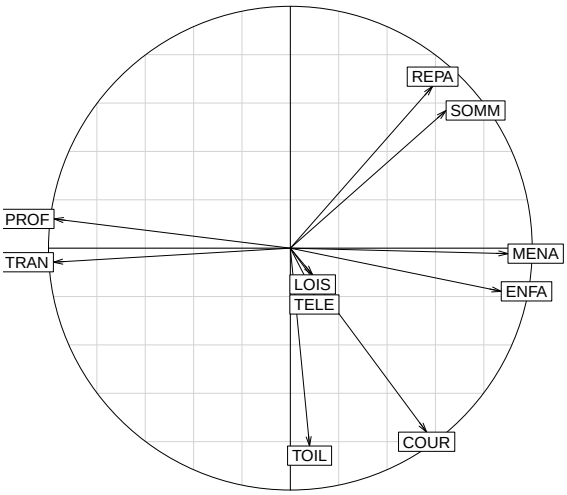
Axe 4

⊖	⊕
HCY (14.2)	FAY (13.6)
HCE (11.5)	FAE (12.4)
	FME (12.1)

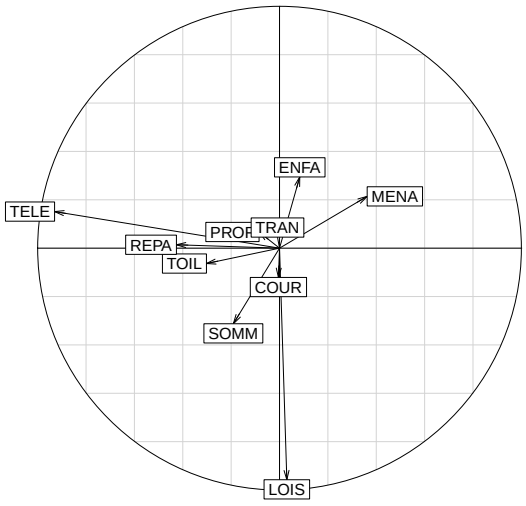
Question 3 *Dessiner les cercles des corrélations sur les deux premiers plans principaux. Quelles sont les variables qui déterminent les axes ?*

On obtient les cercles suivants

```
> s.corcircle(pca1$co)
```




```
> s.corcircle(pca1$co,3,4)
```



Si on se fixe une valeur limite de 0.75 pour la corrélation, les variables qui déterminent les axes sont :

Axe 1

⊖	⊕
PROF $(-0,98)$	MENA $(0,90)$
TRAN $(-0,98)$	ENFA $(0,87)$

Axe 2

⊖	⊕
TOIL $(-0,82)$	
COUR $(-0,76)$	

Axe 3

⊖	⊕
TELE $(-0,93)$	

Axe 4

⊖	⊕
LOIS $(-0,96)$	

Il est intéressant de noter que les loisirs et la télévision ne sont pas des déterminants des différentes catégories sur les deux premiers axes.

Question 4 *En croisant ces résultats avec ceux des individus, donner une nouvelle interprétation des axes.*

— Axe 1 : *Tâches ménagères vs. profession.* les femmes inactives en général et les femmes mariées d'Europe de l'ouest passent plus de temps que la moyenne autour du ménage et de leurs enfants, et bien sûr (pour les inactives) moins de temps au travail et dans les transports ; notons que les hommes non actifs n'apparaissent pas dans nos données. La présence des femmes mariées d'Europe de l'ouest vient probablement du fait qu'elles sont souvent femmes au foyer.
— Axe 2 : *Tâches domestiques non contraintes* : Les femmes américaines passent plus de temps que la moyenne à faire leur toilette et les courses, contrairement aux hommes d'Europe de l'ouest.
— Axe 3 : *Télévision.* les hommes mariés et/ou actifs américains passent beaucoup de temps devant la télé, contrairement aux femmes (sauf célibataires ?) et aux hommes célibataires de Yougoslavie
— Axe 4 : *Loisirs* : il y a une différentiation au sein des pays de l'est, mais elle est peu lisible.

Question 5 *Combien d'axes souhaite-t-on conserver ? Quelle est la qualité globale de représentation dans ces conditions ? Que conclure ?*

Il y a une difficulté en ce qui concerne les axes à conserver, puisque les axes 3 et 4 ne concernent chacun qu'une variable. Deux stratégies sont possibles :
— ne conserver que deux axes, puisque ce sont ceux qui donne une information synthétique entre les variables
— conserver 4 axes, en considérant que les axes 3 et 4 donnent une information sur le fait que télévision et loisir sont décorrélés des autres variables. Les individus auxquels ces variables sont liées ne sont pas si importants.

Je propose ici de conserver 4 axes.

On obtient alors une inertie expliquée égale à $\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 9,23$, soit 92% de l'inertie totale (qui est égale à 10). C'est très bon.