TD1 : Budget-temps (Correction)

1 Budget-temps

1.1 Les données

Il s'agit d'une enquête sur les budgets-temps (temps passé dans différentes activités au cours de la journée). Le tableau suivant comprend 10 variables numériques et 4 variables catégorisées.

```
> require(ade4)
> budget<-read.table("budget.dat")
> budget1<-budget[,1:10]
> pca1<-dudi.pca(budget1, scannf=F, nf=10)
> #
> # seulement utile pour la reproductibilité des TD
> source("fonctions.R")
> pca1 = dudi.fixsigns(pca1, sign.li=c(-1,-1,-1,1,-1))

> budget
```

	PROF	TRAN	MENA	ENFA	COUR	TOIL	REPA	SOMM	TELE	LOIS	SEX	ACT	CIV	PAY
HAU	610	140	60	10	120	95	115	760	175	315	1	1	9	1
FAU	475	90	250	30	140	120	100	775	115	305	2	1	9	1
FNU	10	0	495	110	170	110	130	785	160	430	2	2	9	1
HMU	615	140	65	10	115	90	115	765	180	305	1	9	2	1
FMU	179	29	421	87	161	112	119	776	143	373	2	9	2	1
HCU	585	115	50	0	150	105	100	760	150	385	1	9	1	1
FCU	482	94	196	18	141	130	96	775	132	336	2	9	1	1
HAW	653	100	95	7	57	85	150	808	115	330	1	1	9	2
FAW	511	70	307	30	80	95	142	816	87	262	2	1	9	2
FNW	20	7	568	87	112	90	180	843	125	368	2	2	9	2
HMW	656	97	97	10	52	85	152	808	122	321	1	9	2	2
FMW	168	22	528	69	102	83	174	824	119	311	2	9	2	2
HCW	643	105	72	0	62	77	140	813	100	388	1	9	1	2
FCW	429	34	262	14	92	97	147	849	84	392	2	9	1	2
HAY	650	140	120	15	85	90	105	760	70	365	1	1	9	4
FAY	560	105	375	45	90	90	95	745	60	235	2	1	9	4
FNY	10	10	710	55	145	85	130	815	60	380	2	2	9	4
HMY	650	145	112	15	85	90	105	760	80	358	1	9	2	4
FMY	260	52	576	59	116	85	117	775	65	295	2	9	2	4
HCY	615	125	95	0	115	90	85	760	40	475	1	9	1	4
FCY	433	89	318	23	112	96	102	774	45	408	2	9	1	4
HAE	650	142	122	22	76	94	100	764	96	334	1	1	9	3
FAE	578	106	338	42	106	94	92	752	64	228	2	1	9	3
FNE	24	8	594	72	158	92	128	840	86	398	2	2	9	3
HME	652	133	134	22	68	94	102	763	122	310	1	9	2	3
FME	436	79	433	60	119	90	107	772	73	231	2	9	2	3
HCE	627	148	68	0	88	92	86	770	58	463	1	9	1	3
FCE	434	86	297	21	129	102	94	799	58	380	2	9	1	3

Les 10 variables numériques sont le temps passé en : `PROF`ession, `TRAN`sport, `MENA`ge, `ENFA`nts, `COUR`ses, `TOIL`ette, `REPA`s, `SOMM`eil, `TÉLÉ` et `LOIS`irs.

Les 4 variables catégorisées sont : le `SEX`e (1=Hommes, 2=Femmes), l'`ACT`ivité (1=Actifs, 2=Non Act., 9=Non précisé), l'état `CIV`il (1=Célibataires, 2=Mariés, 9=Non précisé), le `PAY`s (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(ves), `M` : Mariés, `C` : Célibataires, `U` : USA, `W` : Pays de l'Ouest sauf USA, `E` : Est sauf Yougoslavie, `Y` : Yougoslavie.

Les temps sont notés en centièmes d'heures. La première case en haut à gauche du tableau (HAU) indique que les Hommes Actifs des USA passent en moyenne 6 heures et 6 minutes (6 heures + 10/100 d'heure) en activité `PROF`essionnelle. Le total d'une ligne (sur ces 10 variables numériques) est 2400 (24 heures).

Question 1 *Peut-on calculer le temps moyen passé au travail par les hommes aux États-Unis ? Peut-on comparer le temps moyen passé au travail + transport au temps de sommeil pour les hommes américains ?*

La valeur cherchée est une moyenne pondérée de 615 (HMU) et de 595 (HCU). Malheureusement, on ne connaît pas l'effectif de chaque catégorie. Il n'est donc pas possible de calculer la moyenne pour les hommes américains.

Par contre, si on regarde les variables travail+transport (qui s'ajoutent, puisque ce sont des 100e d'heures) et travail pour les hommes américains, on obtient

	PROF+TRAN	SOMM
HMU	755	765
HCU	700	760

Comme les entrées de la dernière colonne sont toutes les deux supérieures à la précédente, quelque soit la pondération employée, la moyenne de PROF+TRAN sera inférieure à celle de SOMM. Le temps de sommeil est donc supérieur au temps de transport+travail pour les hommes américains.

Question 2 *Quelles variables peut-on utiliser pour une analyse en composantes principales ? Les autres apportent-elles une information supplémentaire par rapport au tableau où leur colonnes auraient été retirées ?*

On peut utiliser les variables numériques, c'est-à-dire de PROF à LOIS. Les autres variables sont qualitatives. Elles ne sont pas utiles car :

1. l'information en question est déjà contenue dans les labels des individus
2. On n'a pas toutes les décompositions de chaque variable (on a soit ACT soit CIV).

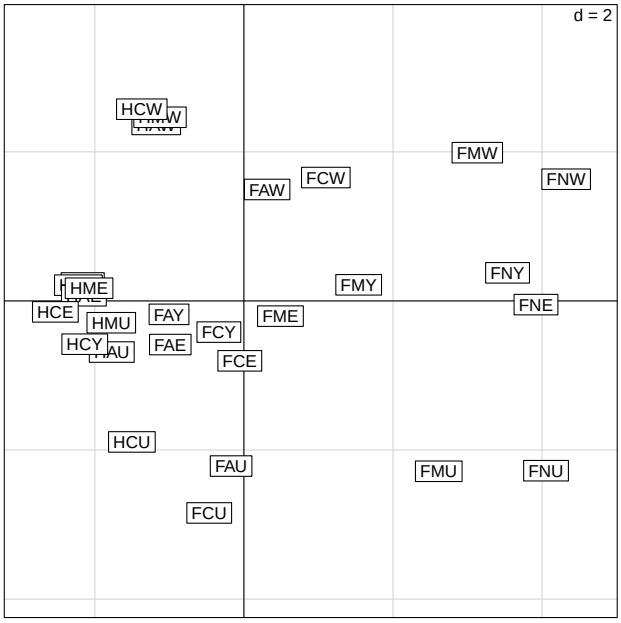
1.2 Analyse en composantes principales

On fait une ACP sur variables centrées-réduites des 10 variables numériques. On donne ci-dessous les coordonnées des individus sur les composantes principales, ainsi que leur représentation sous la forme de la projection sur le premier plan principal (gauche) et sur le plan (3,4) (droite).

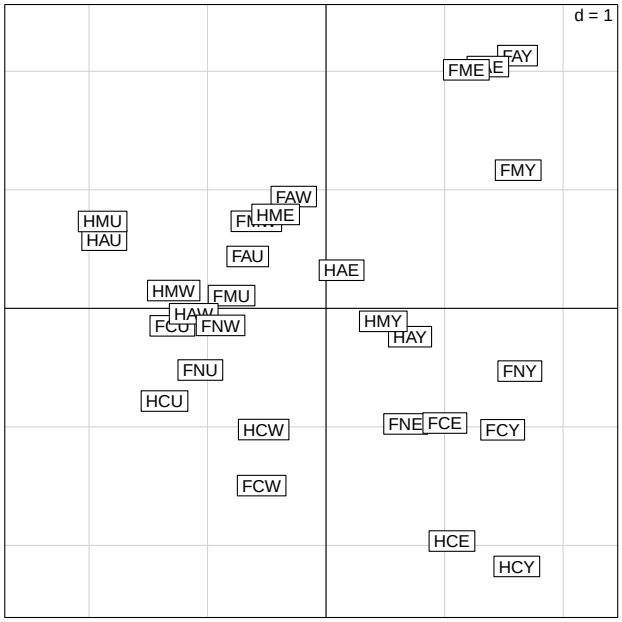
```
> round(pca1$li, 4)
```

```
      Axis1   Axis2   Axis3   Axis4   Axis5   Axis6   Axis7   Axis8   Axis9
HAU -1.7729 -0.6861 -1.8713  0.5752 -0.8544 -0.6656  0.0109 -0.1067  0.1076
FAU -0.1716 -2.2153 -0.6608  0.4376  1.2517 -0.0482  0.0594 -0.0502  0.1340
FNU  4.0534 -2.2777 -1.0605 -0.5203 -1.0370  0.9017 -0.0144  0.2223 -0.1612
HMU -1.7794 -0.2927 -1.8851  0.7330 -1.0384 -0.8302 -0.1849 -0.0843 -0.0050
FMU  2.6143 -2.2853 -0.7972  0.1076 -0.3707  0.5733  0.0614  0.2376 -0.0687
HCU -1.5028 -1.8917 -1.3630 -0.7823 -0.3545 -0.7967  0.3118  0.1619 -0.0546
FCU -0.4652 -2.8443 -1.2964 -0.1476  1.6170  0.1249  0.1035 -0.3277 -0.0055
HAW -1.1763  2.3677 -1.1166 -0.0458  0.2322  0.1413  0.1506  0.0888 -0.0464
FAW  0.3120  1.4953 -0.2724  0.9433  1.2480  0.0129  0.0174  0.0899  0.0957
FNW  4.3234  1.6326 -0.8903 -0.1438 -0.2281  0.4506 -0.0339 -0.2134  0.3458
HMW -1.1254  2.4639 -1.2856  0.1503  0.2178  0.2468  0.1217  0.0801 -0.1320
FMW  3.1313  1.9889 -0.5882  0.7346 -0.3455  0.0113  0.2536 -0.1005  0.1416
HCW -1.3700  2.5720 -0.5263 -1.0228 -0.2872 -0.1318 -0.0847  0.1819 -0.1215
FCW  1.0991  1.6551 -0.5433 -1.4957  1.4300 -0.1123  0.0187  0.2036 -0.2282
HAY -2.1627  0.2410  0.7089 -0.2393 -0.3239  0.3551  0.1249  0.0050  0.2516
FAY -1.0048 -0.1801  1.6157  2.1323 -0.0445  0.2396  0.0954  0.0239 -0.1562
FNY  3.5374  0.3771  1.6353 -0.5295 -0.2760 -0.8007  0.1910 -0.4113 -0.1765
HMY -2.2212  0.2116  0.4832 -0.1096 -0.3972  0.3011  0.0354 -0.0555  0.2595
FMY  1.5400  0.2161  1.6214  1.1658 -0.4390 -0.1844  0.2170 -0.1768 -0.1893
HCY -2.1353 -0.5806  1.6098 -2.1775 -0.5536 -0.0254  0.2554  0.2653  0.0381
FCY -0.3358 -0.4182  1.4906 -1.0249  0.1199  0.1878  0.3154 -0.1425  0.0056
HAE -2.1468  0.0694  0.1312  0.3216 -0.1482  0.5294 -0.3184 -0.1569  0.0237
FAE -0.9877 -0.5854  1.3654  2.0399  0.2673 -0.0864 -0.0199  0.2331  0.0075
FNE  3.9187 -0.0494  0.6719 -0.9754 -0.0023 -0.6430 -0.5950  0.1163  0.0516
HME -2.0774  0.1705 -0.4251  0.7923 -0.2161  0.5265 -0.3649 -0.2804 -0.3499
FME  0.4904 -0.2040  1.1839  2.0125  0.0425 -0.2900 -0.1619  0.3435  0.1267
HCE -2.5294 -0.1468  1.0625 -1.9634 -0.3519  0.3835 -0.2603 -0.1963  0.0396
FCE -0.0551 -0.8035  1.0024 -0.9678  0.8422 -0.3712 -0.3051  0.0491  0.0663
```

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




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



Question 3 *Pour chacune des 4 premières composantes principales, donner la liste des individus qui contribuent à l'axe de manière significative.*

Comme on n'a pas pour l'instant de méthode pour trouver les individus significatifs, on regarde juste les coordonnées les plus grandes en valeur absolue. On fait bien attention de séparer les coordonnées positives des coordonnées négatives.

- axe 1 : négatif `HCE` $(-2,52)$, `HMY` $(-2,22)$, `HAY` $(-2,16)$, `HAE` $(-2,14)$, `HCY` $(-2,13)$, `HME` $(-2,07)$; positif `FNW` $(4,32)$, `FNU` $(4,05)$, `FNE` $(3,91)$, `FNY` $(3,53)$, `FMW` $(3,13)$, `FMU` $(2,61)$;
- axe 2 : négatif `FCU` $(-2,84)$, `FMU` $(-2,28)$, `FNU` $(-2,27)$, `FAU` $(-2,21)$, `HCU` $(-1,89)$, positif `HCW` $(2,57)$, `HMW` $(2,46)$, `HAW` $(2,36)$, `FMW` $(1,98)$, `FCW` $(1,65)$, `FNW` $(1,63)$, `FAW` $(1,49)$;
- axe 3 : négatif `HMU` $(-1,88)$, `HAU` $(-1,87)$, `HCU` $(-1,36)$, `FCU` $(-1,29)$, `HMW` $(-1,28)$, positif `FNY` $(1,63)$, `FMY` $(1,62)$, `FAY` $(1,61)$, `HCY` $(1,61)$, `FCY` $(1,49)$, `FAE` $(1,36)$, `FME` $(1,18)$;
- axe 4 : négatif `HCY` $(-2,18)$, `HCE` $(-1,96)$, `FCW` $(-1,49)$, positif `FAY` $(2,13)$, `FAE` $(2,04)$, `FME` $(2,01)$, `FMY` $(1,16)$.

Question 4 *En utilisant la description des noms des individus donnée en introduction, donner une interprétation des axes (au moins des deux premiers).*

- axe 1 : séparation est/ouest
  - négatif : les hommes de Yougoslavie et autres pays de l'est.
  - positif : les femmes non actives en général et aussi les femmes mariées des US et des autres pays occidentaux.
- axe 2 : séparation USA/reste de l'ouest
  - négatif : les femmes américaines et les hommes célibataires américains.
  - positif : pays de l'ouest.
- axe 3 : ? ? ?
  - négatif : les hommes américains, les femmes célibataires américaines et les hommes mariés de l'ouest.
  - positif : les femmes des pays de l'est et les hommes célibataires de Yougoslavie.
- axe 4 : opposition hommes/femmes à l'est
  - négatif : les hommes célibataires de l'est.
  - positif : les femmes actives et mariées de l'est.

Ces interprétations ne sont pas très précises parce qu'il nous manque l'analyse des variables.

2 Calcul de l'inertie totale

On cherche à établir quelques formules pour l'inertie totale $I_{\mathbf{g}} = \sum_{i=1}^{n} p_i \|\mathbf{e}_i - \mathbf{g}\|_{\mathbf{M}}^2$ d'un nuage de n points de p variables, **sans utiliser la notion de matrice ou de trace de matrice** comme dans le cours.

Question 5 *Si on note σ_j^2 la variance de la variable $\mathbf{x}^j$, montrer que $I_{\mathbf{g}} = \sum_{j=1}^{p} m_j \sigma_j^2$.*

En utilisant la définition de la norme, on peut récrire l'inertie totale comme

$$I_{\mathbf{g}} = \sum_{i=1}^{n} p_i \|\mathbf{e}_i - \mathbf{g}\|^2_{\mathbf{M}} = \sum_{i=1}^{n} p_i \sum_{j=1}^{p} m_j (x_i^j - \bar{x}^j)^2.$$

En échangeant les deux sommes, on obtient

$$I_{\mathbf{g}} = \sum_{j=1}^{p} m_j \sum_{i=1}^{n} p_i (x_i^j - \bar{x}^j)^2 = \sum_{j=1}^{p} m_j \sigma_j^2.$$

Question 6 *Soit $\mathbf{c}'_k = (c_{1k}, \ldots, c_{nk})$ la composante principale (de variance λ_k) associée à l'axe k ; on peut écrire $\mathbf{e}_i - \mathbf{g} = \sum_{k=1}^{p} c_{ik}\mathbf{a}_k$, où les vecteurs $\mathbf{a}_k$ sont $\mathbf{M}$-orthonormés. Montrer que $I_{\mathbf{g}} = \sum_{k=1}^{p} \lambda_k$.*

On procède comme dans la question précédente, mais en représentant chaque individu par ses composantes principales

$$I_{\mathbf{g}} = \sum_{i=1}^{n} p_i \|\mathbf{e}_i - \mathbf{g}\|^2_{\mathbf{M}} = \sum_{i=1}^{n} p_i \sum_{k=1}^{p} \sum_{\ell=1}^{p} c_{ik} c_{i\ell} \langle \mathbf{a}_k, \mathbf{a}_\ell \rangle_{\mathbf{M}} = \sum_{i=1}^{n} p_i \sum_{k=1}^{p} c_{ik}^2.$$

Ici encore, en échangeant les sommes, on écrit

$$I_{\mathbf{g}} = \sum_{k=1}^{p} \sum_{i=1}^{n} p_i c_{ik}^2 = \sum_{k=1}^{p} \operatorname{var} \mathbf{c}_k = \sum_{k=1}^{p} \lambda_k.$$