- Analyse de données -
TP 2 Analyse en composantes principales

---

1 Budget-temps

1. La valeur cherchée est une moyenne pondérée de 615 (HMU) et de 595 (HCU). Malheureusement, on ne connait pas l'effectif de chaque catégorie. Il n'est donc pas possible de calculer la moyenne pour les hommes américains.
2. On peut utiliser les variables numériques, c'est-à-dire de PROF à LOIS. Les autres variables sont qualitatives.
3. 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).
4. ```
   budget<-read.table("budget.txt")
   \# on supprime les 4 dernières variables (qualitatives)
   q=ncol(budget)-4
   budget=budget[,1:q]
   ```
   Puisque on ne retient que 10 variables pour l'ACP, la matrice de corrélations est de dimension (10,10). Il y a donc 10 valeurs propres.
   Tester les deux commandes :
   ```
   acp<-dudi.pca(budget,scann=T)
   acp<-dudi.pca(budget,scann=F)
   ```
   La première permet de visualiser les valeurs propres directement.
   Pour obtenir les valeurs propres :
   ```
   round(acp$eig,4)
   ```
   On sait que la somme des valeurs propres est égale à la dimension de la matrice carré : 10. Puisque la somme des 9 premières valeurs propres est égale à 10 :
   ```
   sum(acp$eig)
   ```
   On en déduit qu'il y a une valeur propre nulle.
   Puisque V admet une valeur propre nulle alors $det(V - 0 \times Id) = det(V) = 0$. Donc les vecteurs colonnes (correspondant aux variables) sont liés. Les variables sont donc liés.
   Retour aux 10 variables : leur somme est égale à 24h (numériquement 2400 ici). La connaissance de 9 variables permet de calculer la 10ème.
5. La règle de Kaiser conduit à retenir les valeurs propres qui sont supérieures à 1, ce qui signifie ici les 4 premières.
6. Les coordonnées des individus sur les premières composantes principales,
   ```
   round(acp$li, 4)
   ```
   ```

	Axis1	Axis2	Axis3	Axis4	Axis5
HAU	-1.7729	-0.6861	-1.8713	0.5752	-0.8544
FAU	-0.1716	-2.2153	-0.6608	0.4376	1.2517
FNU	4.0534	-2.2777	-1.0605	-0.5203	-1.0370
HMU	-1.7794	-0.2927	-1.8851	0.7330	-1.0384
FMU	2.6143	-2.2853	-0.7972	0.1076	-0.3707
HCU	-1.5028	-1.8917	-1.3630	-0.7823	-0.3545
FCU	-0.4652	-2.8443	-1.2964	-0.1476	1.6170
HAW	-1.1763	2.3677	-1.1166	-0.0458	0.2322
FAW	0.3120	1.4953	-0.2724	0.9433	1.2480
FNW	4.3234	1.6326	-0.8903	-0.1438	-0.2281
HMW	-1.1254	2.4639	-1.2856	0.1503	0.2178
FMW	3.1313	1.9889	-0.5882	0.7346	-0.3455
HCW	-1.3700	2.5720	-0.5263	-1.0228	-0.2872
FCW	1.0991	1.6551	-0.5433	-1.4957	1.4300
HAY	-2.1627	0.2410	0.7089	-0.2393	-0.3239
FAY	-1.0048	-0.1801	1.6157	2.1323	-0.0445
FNY	3.5374	0.3771	1.6353	-0.5295	-0.2760
HMY	-2.2212	0.2116	0.4832	-0.1096	-0.3972
FMY	1.5400	0.2161	1.6214	1.1658	-0.4390
HCY	-2.1353	-0.5806	1.6098	-2.1775	-0.5536
FCY	-0.3358	-0.4182	1.4906	-1.0249	0.1199
HAE	-2.1468	0.0694	0.1312	0.3216	-0.1482
FAE	-0.9877	-0.5854	1.3654	2.0399	0.2673
FNE	3.9187	-0.0494	0.6719	-0.9754	-0.0023
HME	-2.0774	0.1705	-0.4251	0.7923	-0.2161
FME	0.4904	-0.2040	1.1839	2.0125	0.0425
HCE	-2.5294	-0.1468	1.0625	-1.9634	-0.3519
FCE	-0.0551	-0.8035	1.0024	-0.9678	0.8422

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), HCY (1.61), FCY (1.49), FAE (1.36), FME (1.18), FAY (1.16) ;

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) .

7. En utilisant la description des noms des individus, une première interprétation des axes :

   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 et les hommes célibataires des pays de l'est.

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.