TP 1 : Analyse en Composantes Principales (ACP)

Zaineb Smida

Le fichier étudié concerne 25 pays de l'Union Européenne (source : Eurostat 2002). Les variables considérées sont les suivantes :

- région (en 3 catégories : 1 = pays de l'est, 2 = pays du sud, 3 = autres)
- espérance de vie à la naissance pour un homme (en années)
- espérance de vie à la naissance pour une femme (en années)
- population (en milliers d'habitants)
- taux d'activité (en pourcentage)% : %population active/population en âge de travailler
- produit intérieur brut par habitant (en standards de pouvoir d'achat)
- taux d'inflation (en pourcentage)
- taux d'emploi (en pourcentage) : actifs occupés/population en âge de travailler
- taux de chômage (en pourcentage): chômeurs/population active
- taux de chômage longue durée (en pourcentage)
- nombre de mariages (pour 1000 personnes)
- nombre d'abonnés aux services de téléphonie mobile (en milliers)
- variable indicatrice qui vaut 1 pour les pays de l'Europe des 15.

Packages nécessaires :

```
library("FactoMineR")
library("ggcorrplot")
```

1 Question

Importer le fichier `pays-eu.txt` sous **R**. Vous pouvez l'appeler `pays`

Conseils

Utiliser la fonction `read.table()`. Lorsque la première ligne du fichier txt contient le nom des variables, il faut ajouter l'option `header = T` pour l'indiquer. Vous pouvez visualiser que l'importation s'est correctement effectuée en faisant `View(pays)`. Il peut être utile d'ajouter des noms de lignes en utilisant la fonction `row.names()`.

Solution

```
pays <- read.table("pays_eu.txt", header = T)
row.names(pays) <- pays$PAYS
pays$REGION <- factor(pays$REGION) #variable qualitative
pays$UE15 <- factor(pays$UE15) #variable qualitative
pays2 <- pays[,3:13] #ici on garde les variables quantitatives
```

2 Question

Réaliser une étude univariée rapide des données à l'aide d'indicateurs numériques et de graphiques. Vous pouvez également représenter les nuages de points entre les paires de variables.

Conseils

Voici quelques fonctions **R** de base permettant de faire de l'analyse univariée :

- `summary()` retourne plusieurs indicateurs comme le minimum, le maximum, la moyenne, la médiane, les quartiles,
- `hist()` prend une variable quantitative en entrée et retourne un histogramme,
- `boxplot()` prend une variable quantitative en entrée et retourne une boîte à moustache,
- `pairs()` prend en entrée plusieurs variables quantitatives (sous forme d'un `data.frame` et retourne toutes les nuages de points possibles entre paires de variables)
- la fonction `plot()` s'utilise principalement sur deux variables quantitatives

Solution

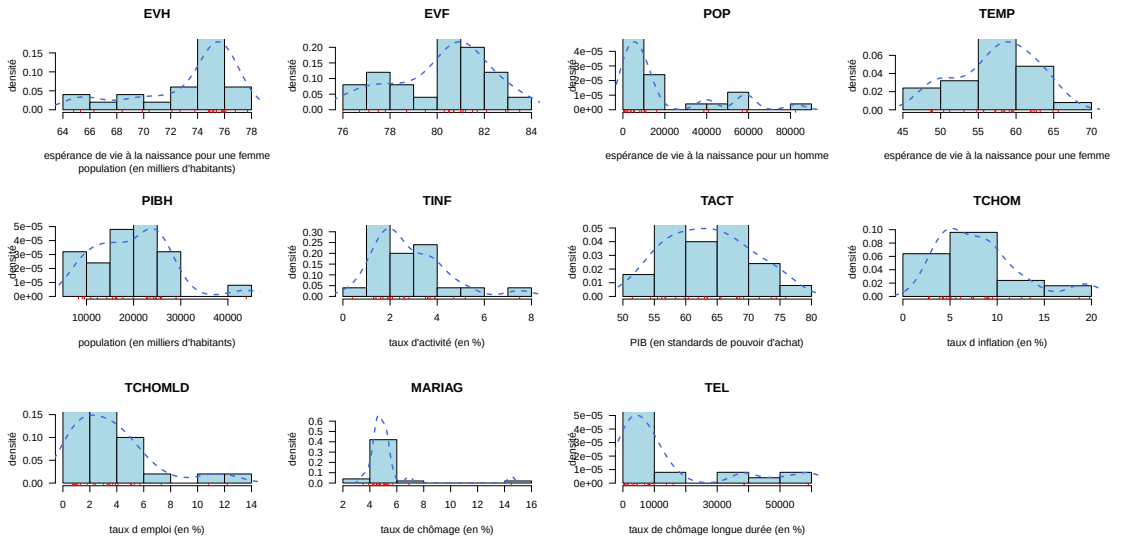
Résumé statistique des variables quantitatives (minium, maximum, moyenne, médiane, quartiles) :

```
summary(pays[, -1]) #on enlève l'identifiant
```

```
 REGION      EVH             EVF             POP               TEMP
 1: 8   Min.   :64.80   Min.   :76.00   Min.   :  394.6   Min.   :48.70
 2: 6   1st Qu.:72.10   1st Qu.:78.70   1st Qu.: 3475.6   1st Qu.:54.80
 3:11   Median :75.10   Median :80.70   Median : 8909.1   Median :58.30
        Mean   :73.41   Mean   :80.23   Mean   :18120.9   Mean   :57.49
        3rd Qu.:75.80   3rd Qu.:81.50   3rd Qu.:16105.3   3rd Qu.:61.90
        Max.   :77.70   Max.   :83.50   Max.   :82440.3   Max.   :65.60
      PIBH            TINF            TACT          TCHOM
 Min.   : 8300   Min.   :0.400   Min.   :51.5   Min.   : 2.700
 1st Qu.:14300   1st Qu.:1.900   1st Qu.:58.4   1st Qu.: 4.900
 Median :19900   Median :2.400   Median :63.4   Median : 7.300
 Mean   :19716   Mean   :2.784   Mean   :63.5   Mean   : 8.096
 3rd Qu.:24600   3rd Qu.:3.600   3rd Qu.:68.2   3rd Qu.: 9.500
 Max.   :43900   Max.   :7.500   Max.   :75.9   Max.   :19.800
    TCHOMLD          MARIAG           TEL        UE15
 Min.   : 0.700   Min.   : 3.50   Min.   :  340   0:10
 1st Qu.: 1.100   1st Qu.: 4.50   1st Qu.: 1632   1:15
 Median : 3.300   Median : 4.70   Median : 7949
 Mean   : 3.612   Mean   : 5.22   Mean   :14209
 3rd Qu.: 5.000   3rd Qu.: 5.20   3rd Qu.:13898
 Max.   :12.200   Max.   :14.50   Max.   :59128
```

Affichage des distributions, en utilisant des histogrammes et densités non paramétriques :

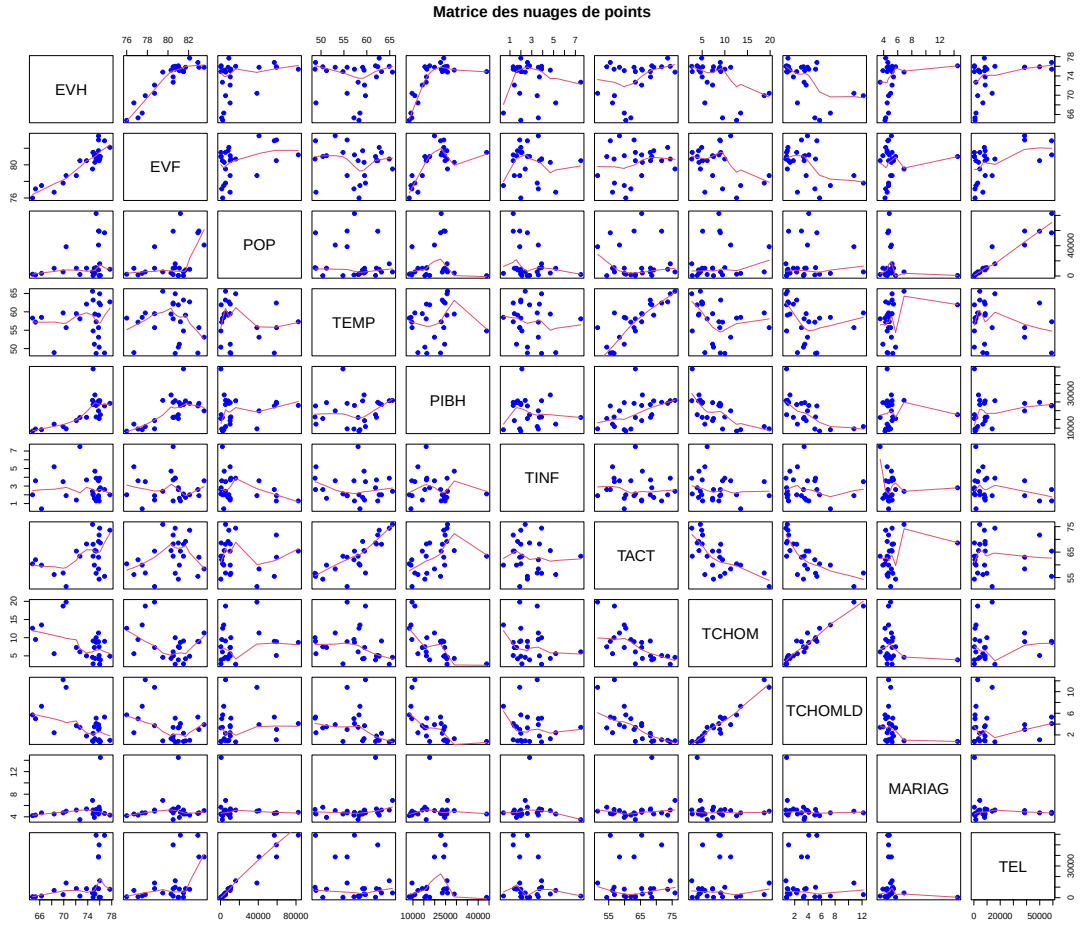
```r
par(mfrow = c(3, 4), las = 1)
nom_var <- c("espérance de vie à la naissance pour un homme",
             "espérance de vie à la naissance pour une femme",
             "population (en milliers d'habitants)",
             "taux d'activité (en %)",
             "PIB (en standards de pouvoir d'achat)",
             "taux d inflation (en %)",
             "taux d emploi (en %)",
             "taux de chômage (en %)",
             "taux de chômage longue durée (en %)",
             "nombre de mariages (pour 1000 personnes)",
             "nombre d'abonnés téléphonie mobile (en milliers)")
for(k in 1:11) {
  temp <- density(pays2[, k])
  hist(pays2[, k], main = names(pays2)[k],
       xlab = nom_var[k-2], ylab = "densité", probability = T,
       col = "lightblue", ylim = range(temp$y))
  lines(temp, col = "royalblue", lty = 2, lwd = 1.5)
  rug(pays2[, k], col = "red")
}
```



Remarque : les ordres de grandeurs sont parfois très différents d'une variable à une autre. En effet, certaines variables sont des comptages (POP, TEL), alors que d'autres sont des pourcentages (TEMP), d'autres des variables quantitatives inférieures à 100 (EVH, EVF); c'est pourquoi, il pourrait être intéressant de reproduire ce graphique sur les données centrées et réduites.

On trace les nuages de points entre chaque paire de variable, ce qui permet de détecter d'éventuels liens (linéaires et non linéaires) entre les paires de variables ainsi que des valeurs atypiques. Par exemple, on constate que le lien entre certaines variables (comme EVH/EVF, TCHOM/TCHOMLD, POP/TEL) est linéaire et très fort. Touefois, il est difficile de lire ce graphique car il contient beaucoup d'informations.

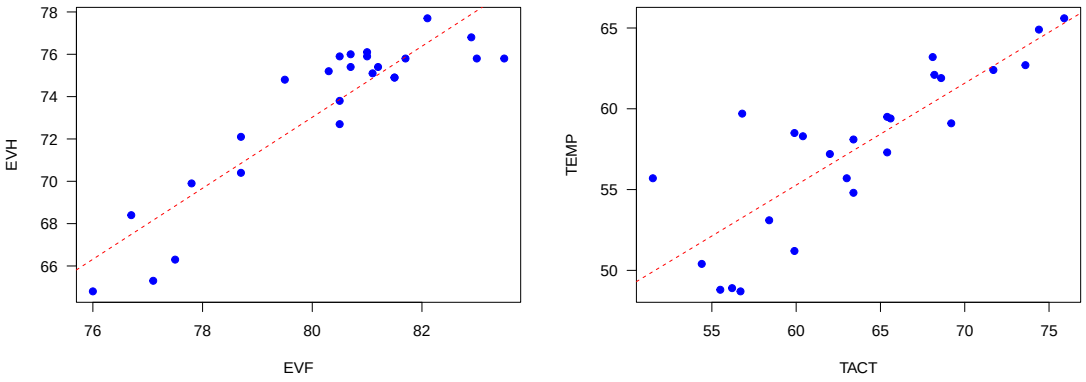
```
pairs(pays2, main = "Matrice des nuages de points",
      pch = 19, col = "blue", panel = panel.smooth)
```



On peut tracer directement les nuages de points qui nous intéressent et représenter par la même occasion la droite de régression linéaire. On représente ici les liens les plus forts.

```r
par(las = 1, mfrow = c(1, 2))
plot(EVH ~ EVF, data = pays2,
     pch = 19,
     col = "blue")
abline(lm(EVH ~ EVF, data = pays2), col = "red", lty = 2)

plot(TEMP ~ TACT, data = pays2,
     pch = 19,
     col = "blue")
abline(lm(TEMP ~ TACT, data = pays2), col = "red", lty = 2)
```



3 Question

Donner la matrice des corrélations et les nuages de points associés. Commenter.

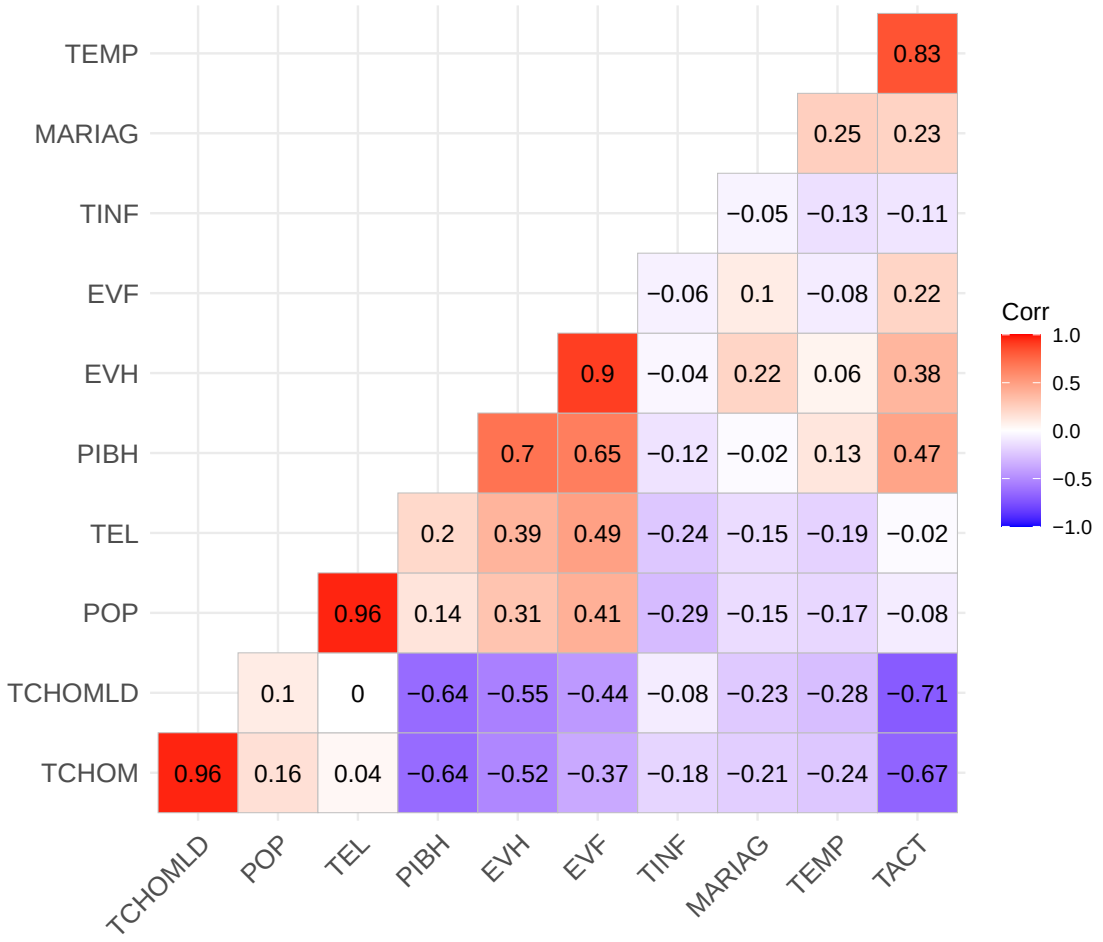
Conseils

- `cor()` prend entrée plusieurs variables quantitatives et retourne la matrice des corrélations,
- la fonction `ggcorrplot()` du package **ggcorrplot** retourne une matrice de corrélation avec une couleur de palette divergente

Solution

On peut représenter la matrice de corrélations linéaires :

```
cor <- cor(pays2)
ggcorrplot(cor, hc.order = TRUE, type = "lower", lab = TRUE)
```



La matrice de corrélations confirme qu'il existe des liens de corrélations entre les variables. Par ailleurs, on constate que certaines variables sont très corrélées entre elles (comme EVH/EVF, TCHOM/TCHOMLD, POP/TEL). Dans ce cas, il peut être improductif d'inclure toutes ces variables dans l'ACP. Une solution consisterait à ne choisir qu'une variable sur les deux.

4 Question

L'ACP vous paraît-elle justifiée ?

Solution

La matrice de corrélations confirme qu'il existe des liens de corrélations entre les variables. Par ailleurs, il est difficile d'interpréter tous les liens qui existent, c'est pourquoi l'ACP semble être appropriée.

5 Question

Déterminer le nombre de composantes principales à retenir pour cette ACP. Justifier votre réponse.

Conseils

- La fonction `PCA()` du package **FactoMineR** permet de réaliser une ACP,
- La fonction `str()` appliquée sur le résultat de l'ACP permet d'identifier les informations retournées. On accède aux éléments de cet objet avec le symbole `$`

Solution

On utilise la fonction `PCA()`:

```
res <- PCA(pays2, graph = FALSE)
```

On extrait les valeurs propres :

```
eigenvalues <- res$eig[, 1]
eigenvalues
```

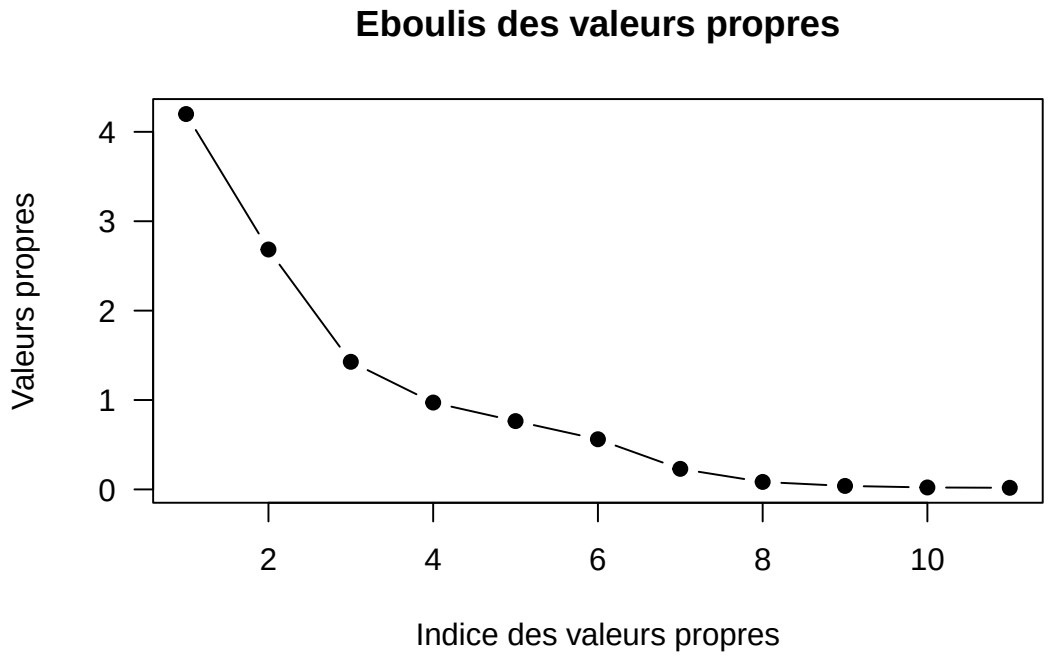
```
    comp 1     comp 2     comp 3     comp 4     comp 5     comp 6     comp 7
4.19930765 2.68386434 1.42753950 0.97162717 0.76418215 0.56105730 0.23004299
    comp 8     comp 9    comp 10    comp 11
0.08363118 0.03890899 0.02187735 0.01796139
```

Selon les critères :

- Critère de Kaiser : 3 vp > 1 mais $\lambda_4 = 0.97$
- Critère de la part d'inertie expliquée : 84% de l'inertie est conservée si l'on retient les 4 premières composantes principales

- Critère de la différence : on va tracer l'éboulis des valeurs propres

```
par(las = 1)
plot(eigenvalues, type = "b", pch = 19,
     xlab = "Indice des valeurs propres",
     ylab = "Valeurs propres",
     main = "Eboulis des valeurs propres")
```



Conclusion : on pourrait choisir de retenir 3 ou 4 axes

6 Question

Interpréter les composantes principales retenues à l'aide des variables initiales. Donner un (des) graphique(s) permettant de visualiser l'interprétation.

Conseils

- On pourra accéder aux coordonnées des variables à partir de l'objet retourné par la fonction `PCA()`. Il pourra être utile de regarder aussi les contributions. La fonction `plot()` appliqué sur les résultats de l'ACP, en ajoutant l'option `choix = "var"` permet de représenter le graphique des variables

Solution

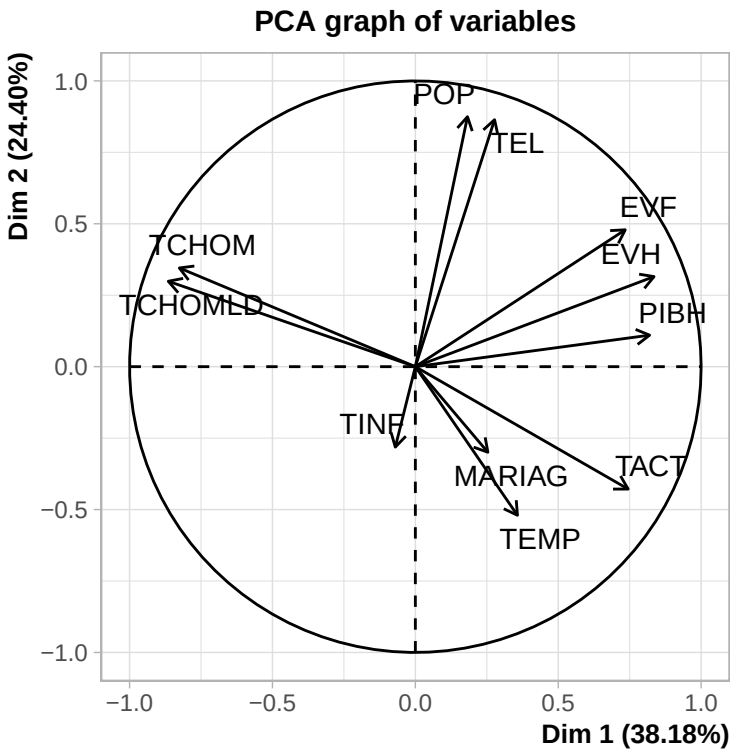
On va d'abord représenter les coordonnées des variables sur les composantes principales :

```
res$var$coord[, 1:4]
```

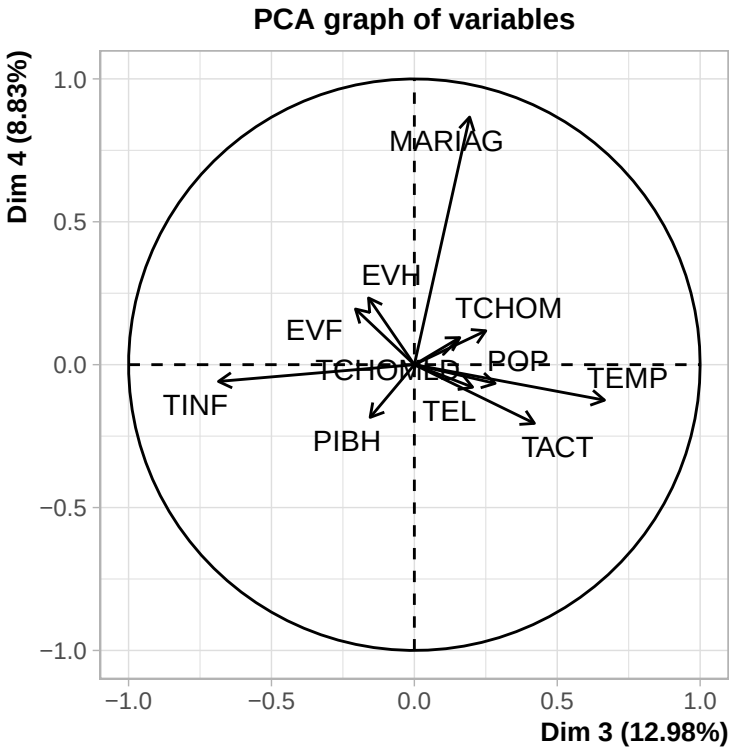
```
             Dim.1      Dim.2      Dim.3       Dim.4
EVH      0.83515555  0.3151215 -0.1602946  0.23336878
EVF      0.73441254  0.4789326 -0.2059038  0.19479526
POP      0.18248329  0.8746272  0.2825919 -0.06550117
TEMP     0.35720093 -0.5199598  0.6653152 -0.12422034
PIBH     0.81964575  0.1102635 -0.1547392 -0.18439581
TINF    -0.07017528 -0.2814160 -0.6853916 -0.05865301
TACT     0.74457183 -0.4282642  0.4201799 -0.20533899
TCHOM   -0.82575580  0.3453544  0.2499975  0.11863522
TCHOMLD -0.86450799  0.2988673  0.1589001  0.09460270
MARIAG   0.25374755 -0.2994415  0.1934134  0.86641148
TEL      0.27713346  0.8641966  0.2037952 -0.07874624
```

On peut représenter le graphique des variables ainsi :

```
plot(res, axes = c(1, 2), choix = "var")
```




```
plot(res, axes = c(3, 4), choix = "var")
```



- L'axe 1 est corrélé positivement à des variables liées au développement d'un pays (espérance de vie H/F, PIB, TACT). Il est négativement corrélée aux deux variables de chômage (TCHOM et TCHOMLD).
- L'axe 2 est corrélé positivement à des variables liées à la population (POP et TEL)
- L'axe 3 oppose clairement la variable taux d'emploi (corrélé positivement à l'axe 3) à l'inflation (corrélée négativement à l'axe 3). La variable TEMP est assi corrélée négativement à l'axe 2, mais comme la contribution de cette variable est plus forte sur l'axe 3, on ne considérera pas cette variable sur l'axe 2.
- L'axe 4 est corrélé positivement à la variable MARIAGE.

On peut aussi regarder les contributions des variables : une forte contribution sera une contribution supérieur à $100/11 = 9$; il semble que nous n'ayons pas omis d'interpréter des variables par rapport aux axes.

```
res$var$contrib[, 1:4]
```

```
            Dim.1      Dim.2      Dim.3      Dim.4
```

```
EVH     16.6095189  3.6999478  1.799906  5.6051323
EVF     12.8440643  8.5464985  2.969891  3.9053243
POP      0.7929915 28.5026585  5.594116  0.4415689
TEMP     3.0384176 10.0734673 31.007504  1.5881291
PIBH    15.9983316  0.4530053  1.677307  3.4994715
TINF     0.1172710  2.9507803 32.907084  0.3540634
TACT    13.2018718  6.8338118 12.367514  4.3395349
TCHOM   16.2377395  4.4439529  4.378074  1.4485305
TCHOMLD 17.7975546  3.3280990  1.768724  0.9211013
MARIAG   1.5332960  3.3408989  2.620506 77.2589391
TEL      1.8289433 27.8268798  2.909375  0.6382048
```

7 Question

Quels sont les pays bien représentés sur chacun des axes retenus ? Justifier votre réponse.

Conseils

- On pourra accéder aux cosinus carré des individus à partir de l'objet retourné par la fonction `PCA()`.

Solution

On récupère les cosinus carrés des individus sur les composantes principales :

```
df <- res$ind$cos2[, 1:4]
df
```

```
               Dim.1        Dim.2        Dim.3        Dim.4
Belgique 0.004237840 0.1088965037 0.2248240027 5.687227e-03
Rep-tche 0.098233170 0.2287960146 0.3620023808 8.574434e-03
Danemark 0.404038614 0.4071776952 0.1590899843 7.132980e-04
Allemagn 0.080645377 0.6794905883 0.1502900132 9.236351e-03
Estonie  0.635206652 0.2112150578 0.0004399479 5.630822e-02
Grece    0.174725892 0.0928860916 0.5424275720 1.154031e-01
Espagne  0.013824844 0.7146677814 0.0866228289 4.322524e-02
France   0.186138919 0.7602915839 0.0129193256 8.918383e-05
Irlande  0.330651372 0.2557234771 0.2455376102 3.336239e-02
Italie   0.008753715 0.9071638310 0.0440483887 9.683098e-03
Chypre   0.146400104 0.1945725913 0.0208837905 6.008295e-01
Lettonie 0.728480936 0.1124378843 0.0692083392 3.658468e-02
```

```
Lituanie 0.684986770 0.0320017502 0.1825275157 1.799627e-03
Luxembou 0.347530374 0.0141998426 0.1309516018 3.380364e-02
Hongrie  0.309962408 0.0477110306 0.3079377448 1.946987e-02
Malte    0.056180050 0.0035394921 0.4270458840 2.006048e-01
Pays-Bas 0.643281575 0.1604836980 0.0142533541 6.275771e-02
Autriche 0.687400262 0.0503106106 0.0009406521 4.498561e-02
Pologne  0.726351365 0.1438102835 0.0458574014 2.765173e-02
Portugal 0.209423668 0.4512611534 0.0026448778 1.130299e-02
Slovenie 0.010147408 0.1382042686 0.4032351492 5.917133e-02
Slovaqui 0.772696040 0.0005157353 0.0250872115 4.339829e-03
Finlande 0.285647746 0.1582238933 0.1406399474 1.315081e-03
Suede    0.668032096 0.0657602494 0.0354333985 4.209247e-02
Royaume- 0.407506606 0.1699173624 0.2614943618 6.756529e-02
```

Sur l'axe 1, les pays avec une valeurs supérieures à 0.5 sont :

```
pays$PAYS[df[, 1] > 0.5]
```

```
[1] "Estonie"  "Lettonie" "Lituanie" "Pays-Bas" "Autriche" "Pologne"  "Slovaqui"
[8] "Suede"
```

Sur l'axe 2, les pays avec une valeur supérieure à 0.25 sont :

```
pays$PAYS[df[, 2] > 0.25]
```

```
[1] "Danemark" "Allemagn" "Espagne"  "France"   "Irlande"  "Italie"   "Portugal"
```

Sur l'axe 3, les pays avec une valeur supérieure à 0.15 sont :

```
pays$PAYS[df[, 3] > 0.15]
```

```
 [1] "Belgique" "Rep-tche" "Danemark" "Allemagn" "Grece"    "Irlande"
 [7] "Lituanie" "Hongrie"  "Malte"    "Slovenie" "Royaume-"
```

Sur l'axe 4, les pays avec une valeur supérieure à 0.1 sont :

```
pays$PAYS[df[, 4] > 0.1]
```

```
[1] "Grece"  "Chypre" "Malte"
```

8 Question

Commenter les contributions des pays aux premiers axes.

Conseils

- On pourra accéder aux contributions des individus à partir de l'objet retourné par la fonction `PCA()`.

Solution

On récupère les contributions des individus sur les composantes principales :

```
df <- res$ind$contrib[, 1:4]
df
```

```
               Dim.1       Dim.2       Dim.3        Dim.4
Belgique  0.01631346  0.65589232  2.54585528  0.094619380
Rep-tche  0.23931102  0.87210771  2.59421235  0.090279275
Danemark  3.88106377  6.11968431  4.49531682  0.029612627
Allemagn  1.21239455 15.98324933  6.64636485  0.600126723
Estonie   6.70649601  3.48917412  0.01366380  2.569392617
Grece     0.98948154  0.82303491  9.03611498  2.824528141
Espagne   0.10808475  8.74229615  1.99216864  1.460560785
France    1.48641477  9.49948673  0.30348150  0.003077988
Irlande   1.77161617  2.14381369  3.86996370  0.772563960
Italie    0.13999982 22.70063329  2.07230989  0.669310234
Chypre    3.72491146  7.74592637  1.56305211 66.069928007
Lettonie 10.84139168  2.61816277  3.02980607  2.353120380
Lituanie  9.44714871  0.69057274  7.40519453  0.107270234
Luxembou  4.65864464  0.29782934  5.16377358  1.958431204
Hongrie   4.07072564  0.98038970 11.89639270  1.105106074
Malte     0.33153248  0.03268150  7.41323991  5.116384933
Pays-Bas  5.68991520  2.22102138  0.37086113  2.399108005
Autriche  3.28478693  0.37616140  0.01322258  0.929071462
Pologne  14.41767478  4.46637870  2.67761091  2.372186560
Portugal  0.61887579  2.08652128  0.02299180  0.144360717
Slovenie  0.11802609  2.51513600 13.79656033  2.974488269
Slovaqui 15.03314095  0.01569947  1.43576378  0.364914954
Finlande  1.02711614  0.89018025  1.48760287  0.020437096
Suede     5.23845724  0.80683859  0.81735071  1.426555135
Royaume-  4.94647643  3.22712794  9.33712517  3.544565243
```

On remarque que Chypre a une forte contribution sur l'axe 4.

9 Question

Réaliser le(s) graphique(s) des pays et commenter l'ACP.

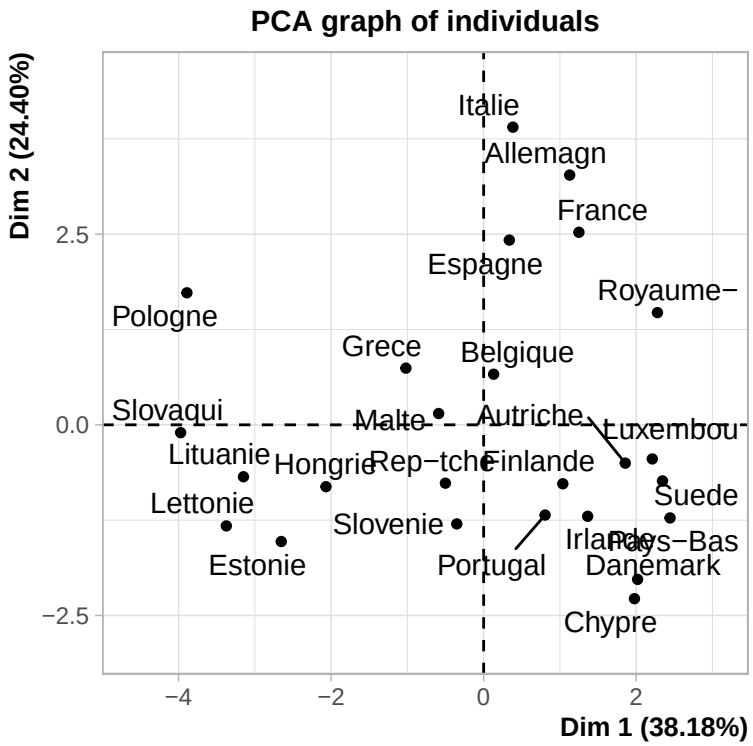
Conseils

- On pourra accéder aux coordonnées des individus à partir de l'objet retourné par la fonction `PCA()`. Il pourra être utile de regarder aussi les contributions. La fonction `plot()` appliqué sur les résultats de l'ACP, en ajoutant l'option `choix = "ind"` permet de représenter le graphique des variables

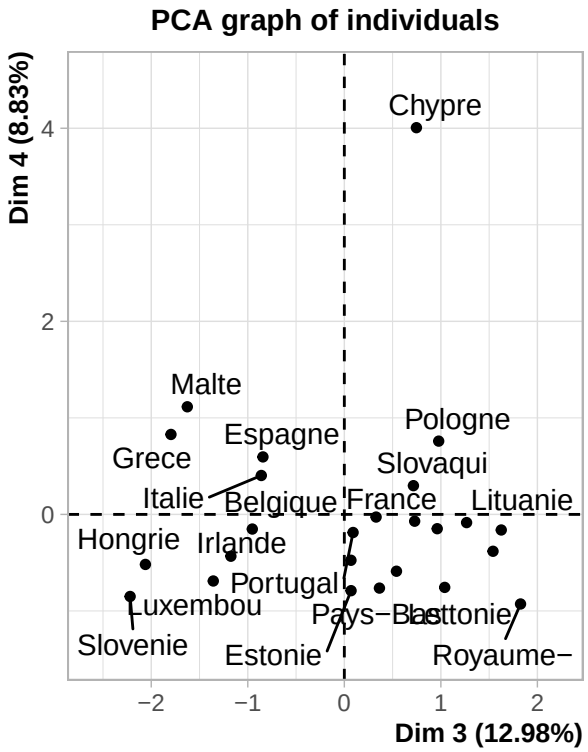
Solution

On peut représenter le graphique des variables ainsi :

```
plot(res, axes = c(1, 2), choix = "ind")
```



```
plot(res, axes = c(3, 4), choix = "ind")
```



- Sur l'axe 1 : les pays comme la Suède, le Pays-Bas et l'Autriche ont des valeurs élevées de EVH, EVH, PIBH et TACT et faibles pour les variables TCHOM et TCHOMLD, alors que cela est plutôt l'inverse pour la Pologne, la Slovaquie, la lettonie, la Lituanie et l'Estonie.
- Sur l'axe 2 : les pays comme la France, l'Allemagne, l'Espagne et l'Italie ont des valeurs de POP et TEL élevés alors que les pays comme le Danemark, l'Irlande et le Portugal ont des valeurs de POP et TEL faibles
- Sur l'axe 3 : les pays comme la République Tchèque, le Danemark, l'Allemagne, la Lituanie, et le Royaume-Uni, ont des valeurs fortes de TEMP et faibles de TINF alors que c'est l'inverse pour la Belgique, la Grèce, l'Irlande, la Hongrie, Malte et la Slovénie.
- Sur l'axe 4 : CHYPRE a un Taux de mariage et dans une moindre mesure, Malte et la Grèce

10 Question

Refaire l'analyse en utilisant la fonction `PCAshiny()` du package **Factoshiny**:

Solution

```
library(Factoshiny)
PCAshiny(pays2)
```