Fiche TD avec le logiciel R : tdr53

Musique

D. Chessel & A.B. Dufour

Exercices simples pour repérer que dans une analyse en composantes principales, il est question en même temps de ressemblances (corrélation entre variables) et de différences (distance entre individus).

Table des matières

1 Introduction

On a demandé à n= 30 étudiants de la filière biomathématique de ranger par ordre de préférence 8 groupes de musique : la personne 1 (première ligne) préfère le 7, ensuite le 8, ensuite le 6, ensuite le 5, ... enfin le 2. Le code des groupes est :

- 1 U2 (`http://www.atu2.com/`)
- 2 ABBA (`http://www.abbasite.com/`)
- 3 Hendrix (`http://www.jimi-hendrix.com/`)
- 4 Les Chaussettes Noires (`http://www.cybercd.fr/artist/Chaussettes+Noires`)
- 5 Zappa (`http://www.zappa.com/`)
- 6 Doors (`http://www.3doorsdown.com/`)
- 7 Bob Marley (`http://www.bobmarley.com/`)
- 8 Léo Ferré (`http://www.leoferre.org/`)

Les données sont dans le fichier `u2toferre.txt`.

```
ordi <- read.table(url("http://pbil.univ-lyon1.fr/R/donnees/u2toferre.txt"))
head(ordi)
```

```
  V1 V2 V3 V4 V5 V6 V7 V8
1  7  8  6  5  1  3  4  2
2  1  2  6  7  3  8  4  5
3  6  3  2  1  7  8  5  4
4  3  6  5  7  1  8  4  2
5  3  1  6  7  5  4  8  2
6  8  7  4  5  6  3  2  1
```

2 Identiques et opposés

Y-a-t-il deux étudiants qui ont fait exactement le même choix ? Sinon, quels sont les choix les plus voisins ? Y-a-t-il deux étudiants qui ont fait exactement un choix inverse ? Sinon, quels sont les choix les plus différents ?

`ordi` ressemble à un tableau de données mais n'est pas un tableau de données !

```
apply(ordi, 1, order)[, 1:20]
```

```
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14]
[1,]    5    1    4    5    2    8    1    3    1     1     1     2     1     3
[2,]    8    2    3    8    8    7    7    4    3     5     4     5     5     7
[3,]    6    5    2    1    1    6    3    6    2     4     5     3     4     4
[4,]    7    7    8    7    6    3    6    8    7     7     7     8     6     8
[5,]    4    8    7    3    5    4    4    7    8     6     6     6     7     5
[6,]    3    3    1    2    3    5    5    2    4     2     2     1     2     1
[7,]    1    4    5    4    4    2    8    1    6     3     3     4     3     2
[8,]    2    6    6    6    7    1    2    5    5     8     8     7     8     6
     [,15] [,16] [,17] [,18] [,19] [,20]
[1,]     3     1     3     1     3     1
[2,]     1     4     1     6     8     4
[3,]     4     2     4     2     7     3
[4,]     6     6     7     8     6     6
[5,]     8     8     6     7     4     7
[6,]     5     3     2     3     5     2
[7,]     2     7     5     5     2     5
[8,]     7     5     8     4     1     8
```

On a ici le rang du groupe dans un choix individuel.

```
rang <- as.data.frame(t(apply(ordi, 1, order)))
names(rang) <- c("U2", "ABBA", "HENDRIX", "Chau_Noir", "ZAPPA",
    "DOORS", "MARLEY", "FERRE")
```

Editer le tableau et ses sommes marginales.

```
apply(rang, 1, sum)
```

```
 [1] 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36
[28] 36 36 36
```

```
apply(rang, 2, sum)
```

```
      U2      ABBA    HENDRIX Chau_Noir     ZAPPA     DOORS    MARLEY     FERRE
      77       149       113       203       185        94        96       163
```

On s'intéresse à la différence qui existe entre deux étudiants. Prenons par exemple les deux premiers.

```
sqrt(sum((rang[1, ] - rang[2, ])^2))
```

```
[1] 9.69536
```

Prenons les 6 premiers :

```
dist(rang[1:6, ])
```

```
          1         2         3         4         5
2  9.695360
3  9.380832  5.099020
4  7.211103  9.695360  6.782330
5  8.366600  8.000000  6.633250  4.000000
6  5.656854 11.832160 11.489125  9.486833 10.770330
```

Prenons les tous :

```
dpetite <- dist(rang[1:6, ])
d0 <- dist(rang)
```

La distance entre deux personnes exprime la dissemblance de leur choix. Étudier l'objet `d0`. Quelle est sa classe ? Est-ce une liste (`is.list`), un vecteur (`is.vector`), une matrice (`is.matrix`), un data.frame (`is.data.frame`), un facteur (`is.factor`), un numérique (`is.numeric`) ? Regarder sa forme exacte avec `unclass` :

```
unclass(dpetite)
```

```
 [1]  9.695360  9.380832  7.211103  8.366600  5.656854  5.099020  9.695360  8.000000
 [9] 11.832160  6.782330  6.633250 11.489125  4.000000  9.486833 10.770330
attr(,"Size")
[1] 6
attr(,"Labels")
[1] "1" "2" "3" "4" "5" "6"
attr(,"Diag")
[1] FALSE
attr(,"Upper")
[1] FALSE
attr(,"method")
[1] "euclidean"
attr(,"call")
dist(x = rang[1:6, ])
```

Observer que ces objets ont des attributs :

```
attributes(d0)
```

```
$Size
[1] 30
$Diag
[1] FALSE

$Upper
[1] FALSE

$method
[1] "euclidean"

$call
dist(x = rang)

$class
[1] "dist"
```

Expliquer enfin la nature exacte de cet objet. Si deux choix sont identiques, la distance est nulle.

```
 min(d0)
```

```
[1] 1.414214
```

```
 d0.mat <- as.matrix(d0)
 min(d0.mat)
```

```
[1] 0
```

Expliquer le pourquoi de cette édition. On cherche les jugements les plus proches les uns des autres :

```
 (1:900)[d0.mat == min(d0)]
```

```
[1] 281 283 310 370
```

Il y a 4 valeurs de distances minimales (2 couples). Où sont-ils ?

```
 col(d0.mat)[d0.mat == min(d0)]
```

```
[1] 10 10 11 13
```

```
 row(d0.mat)[d0.mat == min(d0)]
```

```
[1] 11 13 10 10
```

```
 rang[c(10, 11, 13), ]
```

```
   U2 ABBA HENDRIX Chau_Noir ZAPPA DOORS MARLEY FERRE
10  1    5       4         7     6     2      3     8
11  1    4       5         7     6     2      3     8
13  1    5       4         6     7     2      3     8
```

Ils ont des goûts très voisins ! Continuer :

```
 rang[c(6, 17, 20), ]
```

```
   U2 ABBA HENDRIX Chau_Noir ZAPPA DOORS MARLEY FERRE
6   8    7       6         3     4     5      2     1
17  3    1       4         7     6     2      5     8
20  1    4       3         6     7     2      5     8
```

Ils ont des goûts très dissemblables.

3 Représenter les différences

On voudrait positionner sur un axe les n personnes de manière à ce que la distance entre deux points sur cet axe représente au mieux la distance entre les deux points dans la matrice de distances.

Lire la documentation de `cmdscale`.

Faire l'exemple proposé. Lister la fonction et décrire son contenu.

Si $\mathbf{D}$ est la matrice de distance, elle diagonalise $\mathbf{\Delta} = -\frac{1}{2}\left[d_{ij}^2\right]_{\bullet\bullet}$ où le double point indique le double centrage. On démontre [1] que s'il existe un nuage de n points dans un espace euclidien dont les distances deux à deux au carré sont les d_{ij}^2, alors la matrice Δ est celle des produits scalaires entre les points. L'existence de ce nuage est garantie si et seulement si les valeurs propres de Δ sont toutes positives ou nulles. C'est faux pour eurodist :

```
 cmdscale(eurodist, 20, eig = T)$eig
```

```
 [1]  1.953838e+07  1.185656e+07  1.528844e+06  1.118742e+06  7.893472e+05
 [6]  5.816552e+05  2.623192e+05  1.925976e+05  1.450845e+05  1.079673e+05
[11]  5.139484e+04 -3.259629e-09 -9.496124e+03 -5.305820e+04 -1.322166e+05
[16] -2.573360e+05 -3.326719e+05 -5.162523e+05 -9.191491e+05 -1.006504e+06
```

[1] 19538377.090 11856555.334 1528844.468 1118741.951 789347.203 581655.207 [7] 262319.208 192597.562 145084.535 107967.307 51394.841 0.000 [13] -9496.124 -53058.196 -132216.575 -257336.026 -332671.901 -516252.254 [19] -919149.098 -1006503.960 Warning message : NaNs produced in : sqrt(ev) C'est vrai pour d0 (avec 13 chiffres significatifs, merci aux numériciens de R) :

```
cmdscale(d0, 29, eig = T)$eig
```

```
 [1]  2.996247e+02  1.327764e+02  1.181948e+02  7.550933e+01  5.577551e+01
 [6]  4.576754e+01  2.721834e+01  3.581377e-14  9.119011e-15  8.417171e-15
[11]  7.308419e-15  7.020989e-15  5.095523e-15  4.153043e-15  1.812578e-15
[16]  1.149842e-15  4.901473e-16  4.519406e-16 -1.250024e-15 -1.896716e-15
[21] -1.965749e-15 -2.355146e-15 -3.018505e-15 -3.750136e-15 -3.773127e-15
[26] -3.829297e-15 -5.010308e-15 -8.176739e-15 -1.196522e-14
```

C'est normal, puisque les distances ont été calculées à partir d'un nuage euclidien (les lignes du tableau `rang` ou du tableau centré associé, ce qui ne change rien aux distances) :

```
w <- as.matrix(scale(rang, center = T, scale = F))
eigen(t(w) %*% w)$values
```

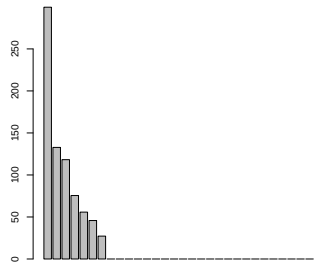
```
[1]  2.996247e+02  1.327764e+02  1.181948e+02  7.550933e+01  5.577551e+01
[6]  4.576754e+01  2.721834e+01 -1.574795e-14
```

```
eigen(w %*% t(w))$values
```

```
 [1]  2.996247e+02  1.327764e+02  1.181948e+02  7.550933e+01  5.577551e+01
 [6]  4.576754e+01  2.721834e+01  1.444699e-14  1.258869e-14  1.094308e-14
[11]  4.650304e-15  2.164563e-15  1.012217e-15  9.816014e-16  6.123915e-16
[16]  4.931675e-16  2.568662e-16  2.174631e-16 -2.218012e-16 -2.447902e-16
[21] -4.979176e-16 -5.755208e-16 -6.991930e-16 -2.622113e-15 -2.682942e-15
[26] -2.866674e-15 -3.432789e-15 -3.554613e-15 -6.327721e-15 -4.571905e-14
```

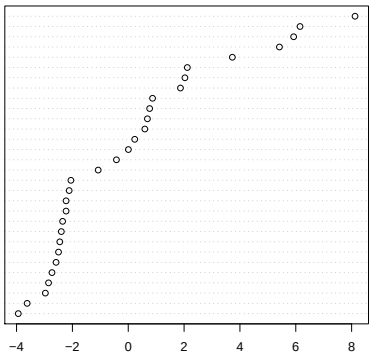
Les valeurs propres non nulles de $\mathbf{X}_0\mathbf{X}_0^T$ sont celles de $\mathbf{X}_0^T\mathbf{X}_0$ et celles de $\Delta = -\frac{1}{2}\left[d_{ij}^2\right]_{\bullet\bullet}$.

```
barplot(eigen(w %*% t(w))$values)
```



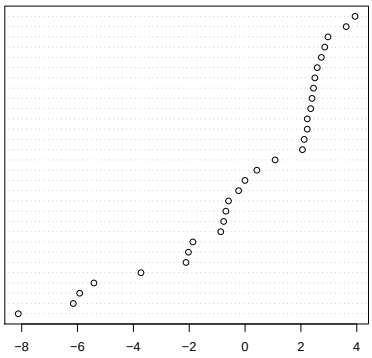
Le positionnement multiple classique (*Classical Multidimensional Scaling*) dans notre cas revient exactement à l'analyse en composantes principales centrée. Pour représenter les distances :

```
w <- cmdscale(d0)[, 1]
dotchart(sort(w))
```



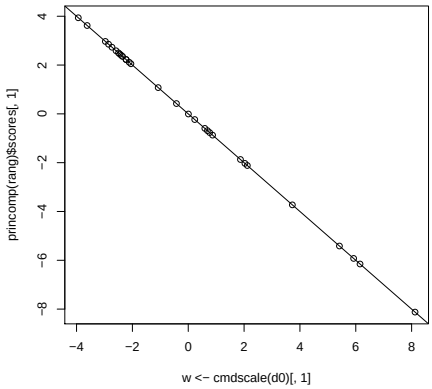
Observer les positions extrêmes (6 contre 20 et 17).

```
dotchart(sort(princomp(rang)$scores[, 1]))
```



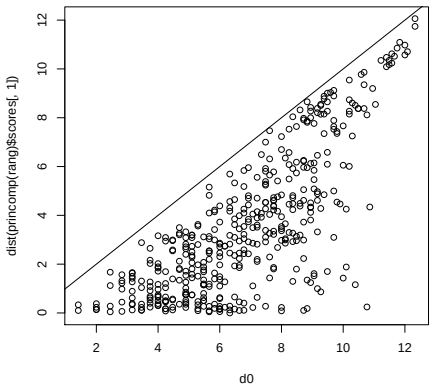
Vérifier qu'au signe près, c'est la même chose.

```
plot(w <- cmdscale(d0)[, 1], princomp(rang)$scores[, 1])
abline(0, -1)
```



On peut alors représenter les distances réelles et celle qui sont exprimées sur le premier axe de l'ordination :

```
plot(d0, dist(princomp(rang)$scores[, 1]))
abline(0, 1)
```



Le graphique est proposé p. 118 dans [2].

```
sum(dist(princomp(rang)$scores[, 1])^2)/sum(d0^2)
```

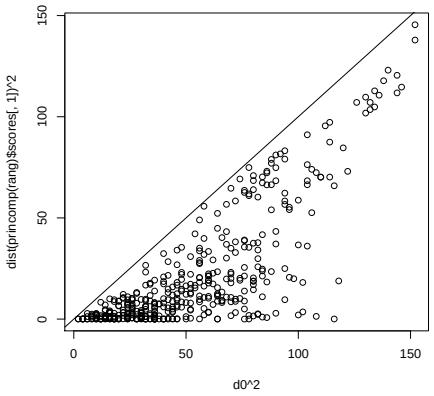
```
[1] 0.396924
```

```
lambda <- princomp(rang)$sdev^2
lambda[1]/sum(lambda)
```

```
  Comp.1
0.396924
```

Le taux d'inertie est de ce point de vue un pourcentage de distance totale représentée. Il vaut donc mieux utiliser le graphe :

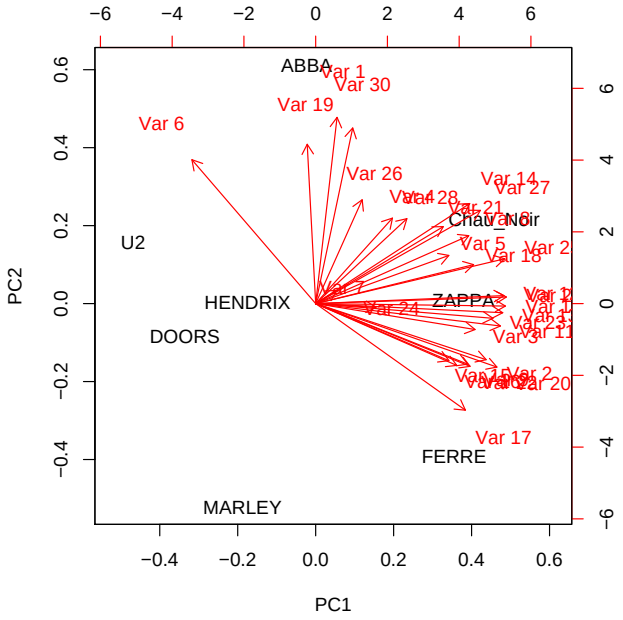
```
plot(d0^2, dist(princomp(rang)$scores[, 1])^2)
abline(0, 1)
```



4 Représenter les ressemblances

L'ACP du tableau de rang où les éléments de choix définissent les variables est donc centrée sur l'analyse des différences. Elle s'oppose totalement à la recherche des éléments de convergence entre les sélections faites. Pour représenter le compromis :

```
z <- t(rang)
biplot(prcomp(z))
```



Expliciter la position des étudiants par rapport à celles des groupes de musique. Retourner aux données :

```
rang[c(1, 19, 30), ]
   U2 ABBA HENDRIX Chau_Noir ZAPPA DOORS MARLEY FERRE
1   5    8       6         7     4     3      1     2
19  3    8       7         6     4     5      2     1
30  5    8       4         6     7     3      2     1
```

5 Cohérences des juges

On considère un jury de dégustation de L juges qui ont à classer C produits pour un critère donné. Chacun des juges classe les produits par ordre de préférence (sans ex æquo) et attribue à chacun un entier compris entre 1 et C : B_{ij} est le rang proposé par le juge i pour le produit j. Le résultat de la dégustation est consigné dans un tableau de L lignes et C colonnes où chaque ligne est une permutation des entiers 1, ..., C. On note m_i et s_i^2 la moyenne et la variance des rangs attribués par le juge i. De plus S_j est la somme des rangs associés au produit j et on pose :

$$T_{ik} = \sum_{j=1}^{c} B_{ij} B_{kj}$$

a) Calculerm_i et s_i^2. b) Calculer le coefficient de corrélation linéaire r_{ik} entre les juges i et k (coefficient de Spearman) en fonction de C et T_{ik}. c) On pose $U_{ik} = \sum_{j=1}^{c} (B_{ij} - B_{kj})^2$. Quelle relation existe-t-il entre r_{ik} et U_{ik} ? d) Calculer la moyenne M et la variance V de la statistique $(S_j)_{j=1,...,C}$ en fonction de L, C et la somme pour $i \neq k$ des T_{ik}. e) Exprimer en fonction de L, C, V la moyenne Z des coefficients de Spearman pour tous les couples (i,k) de juges. f) Vérifier le résultat obtenu lorsque tous les jugements concordent. g) Pour quelle situation la variance V sera-t-elle maximale ? Quelle est alors la variance V_{max} ? h) On pose $K = \frac{V}{V_{max}}$. K est appelé coefficient de concordance de Kendall. Exprimer Z en fonction de K et K en fonction de Z. Justifier le nom donné à K. i) Illustrer numériquement ces résultats sur le tableau de données. j) Illustrer numériquement ce résultats sur un tableau aléatoire :

```
alea <- matrix(unlist(tapply(1:20, 1:20, function(x) sample(1:8,
    8, rep = F))), 20, 8, byr = T)
```

Références

[1] J.C. Gower. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53 :325–338, 1966.

[2] L. Legendre and P. Legendre. *Ecologie numérique, tome 2 : La structure des données écologiques*. Masson, Paris, 1984.