Fiche TD avec le logiciel : tdr24

Axes principaux

D. Chessel, A.B. Dufour & J.R. Lobry

Approches de la définition en dimensions réduite

Table des matières

1 Deux variables centrées **1**

2 Données triangulaires **6**

3 Les axes principaux d'une loi normale **9**

4 Variables non centrées **11**

5 Mode Q et R **12**

6 Valeurs propres doubles **14**
6.1 Imprécision des calculs numériques 14
6.2 Certificat d'études 15

7 Exercices **15**

1 Deux variables centrées

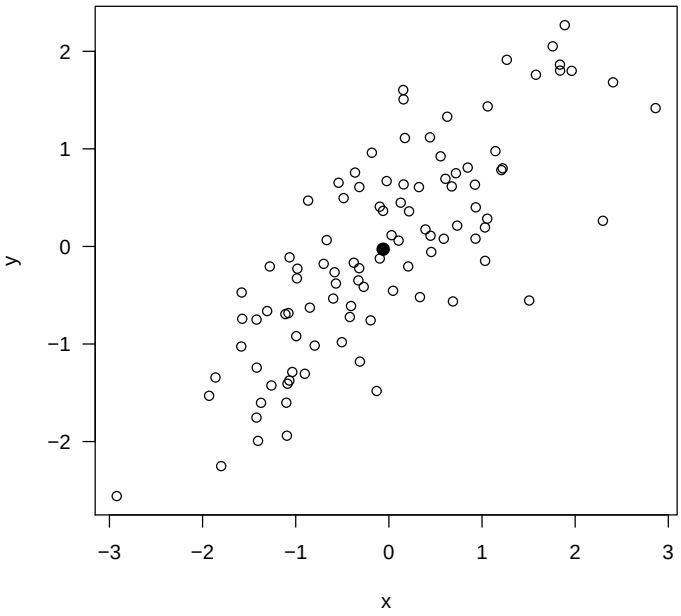
On utilise la fonction `mvrnorm()` de la bibliothèque `MASS`, consulter sa documentation (`?mvrnorm`) :

```
mvrnorm                  package:MASS                  R Documentation

Simulate from a Multivariate Normal Distribution

Description:

     Produces one or more samples from the specified multivariate
     normal distribution.
```

```
library(MASS)
w <- mvrnorm(n = 100, mu = c(0, 0), Sig = matrix(c(1, 0.75, 0.75,
    1), 2, 2))
w <- data.frame(w)
names(w) <- c("x", "y")
```

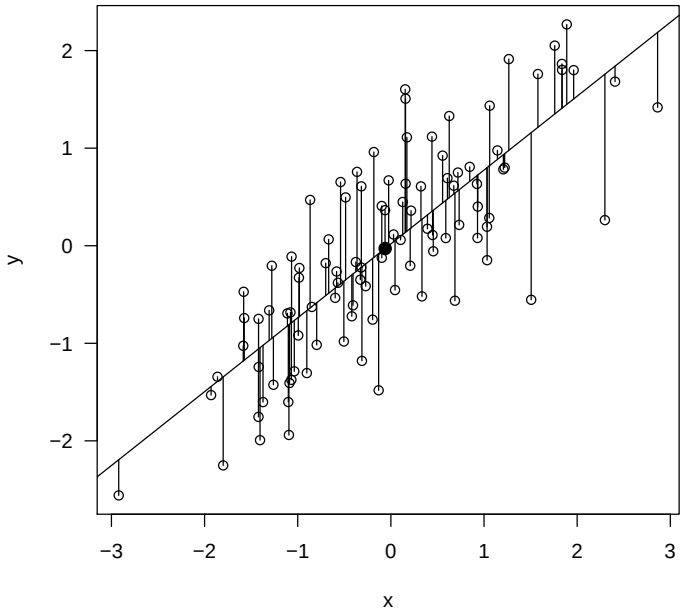
Représenter le nuage de points et sa moyenne :

```
plot(w, las = 1)
points(mean(w$x), mean(w$y), pch = 20, cex = 2)
```



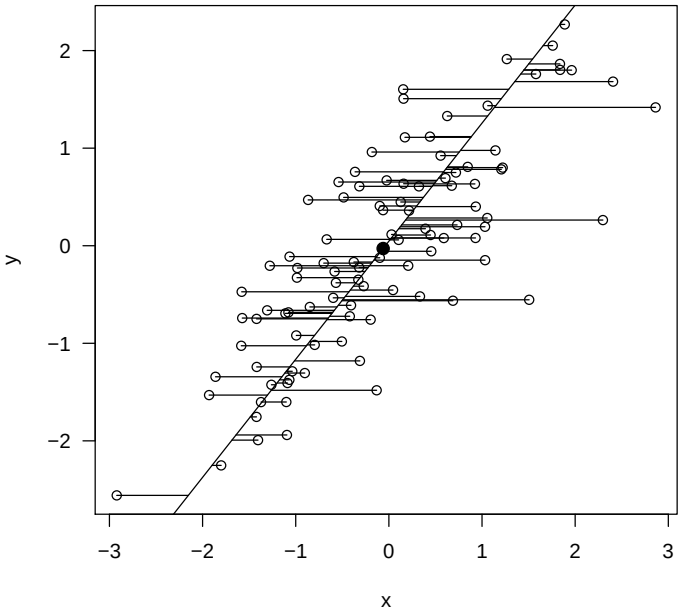
Représenter la droite de régression de y/x et son principe :

```
plot(w, las = 1)
points(mean(w$x), mean(w$y), pch = 20, cex = 2)
abline(lm(w$y ~ w$x))
segments(w$x, w$y, w$x, predict(lm(w$y ~ w$x)))
```



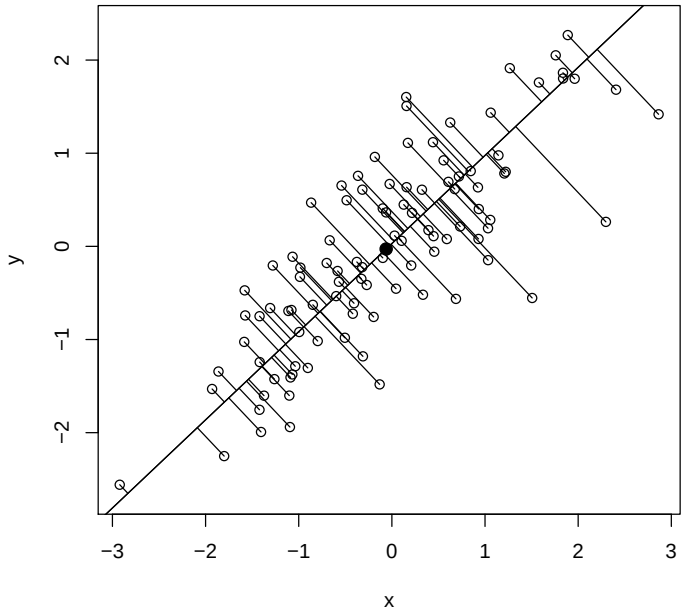
Représenter la droite de régression de x/y et son principe :

```
plot(w, las = 1)
points(mean(w$x), mean(w$y), pch = 20, cex = 2)
a0 <- coefficients(lm(w$x ~ w$y))
abline(-a0[1]/a0[2], 1/a0[2])
segments(w$x, w$y, predict(lm(w$x ~ w$y)), w$y)
```



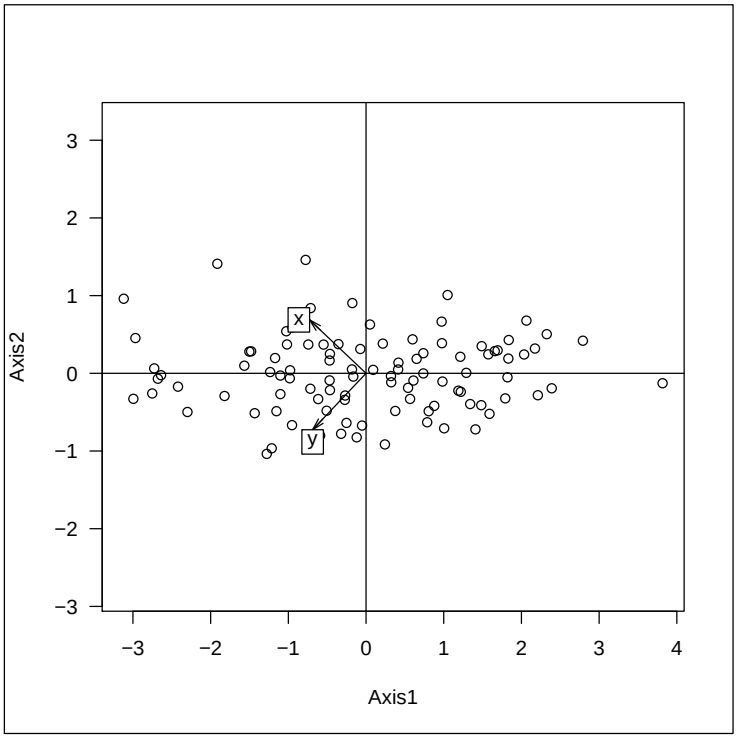
Représenter l'axe principal et son principe :

```
plot(w, asp = 1)
points(mean(w$x), mean(w$y), pch = 20, cex = 2)
cov <- var(w)
u <- eigen(cov, sym = T)$vectors[, 1]
p <- u[2]/u[1]
abline(c(mean(w$y) - p * mean(w$x), p))
scal <- (w$x - mean(w$x)) * u[1] + (w$y - mean(w$y)) * u[2]
abline(c(mean(w$y) - p * mean(w$x), p))
segments(w$x, w$y, mean(w$x) + scal * u[1], mean(w$y) + scal * u[2])
```



On a représenté l'axe principal dans la base canonique. Représenter la base canonique dans la base des axes principaux :

```
library(ade4)
pca1 <- dudi.pca(w, scal = FALSE, scann = FALSE, nf = 2)
plot(pca1$li, asp = 1, las = 1)
abline(h = 0)
abline(v = 0)
s.arrow(pca1$c1, add.p = TRUE)
```



Source : Pearson, K. (1901) On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, **2** :559-572. `http://pbil.univ-lyon1.fr/R/pearson1901.pdf`

2 Données triangulaires

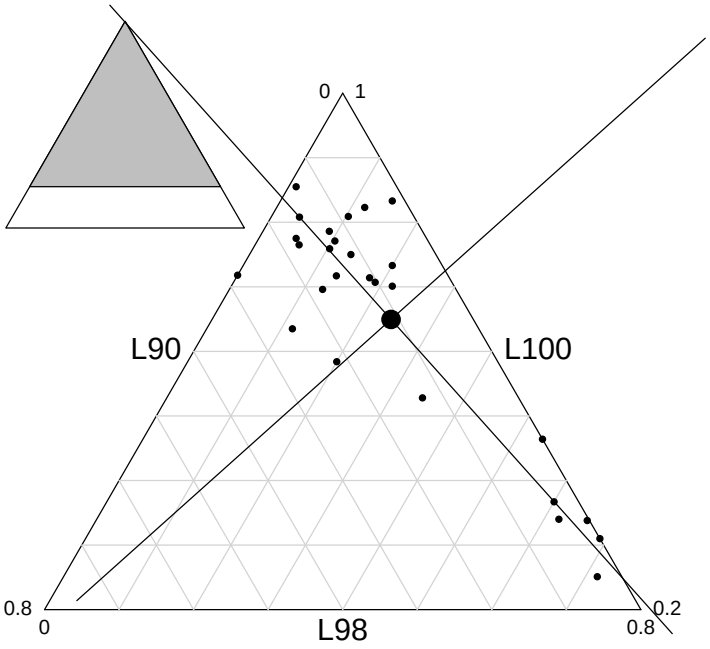
```
locus <- read.table("http://pbil.univ-lyon1.fr/R/donnees/locus.txt",
    header = TRUE, row.names = 1)
locus[1:5, ]
```

```
   L90  L98 L100
1  1.7 15.0 83.3
2  8.3 21.6 70.1
3 10.7 17.9 71.4
4  6.7 20.0 73.3
5 10.3 19.0 70.7
```

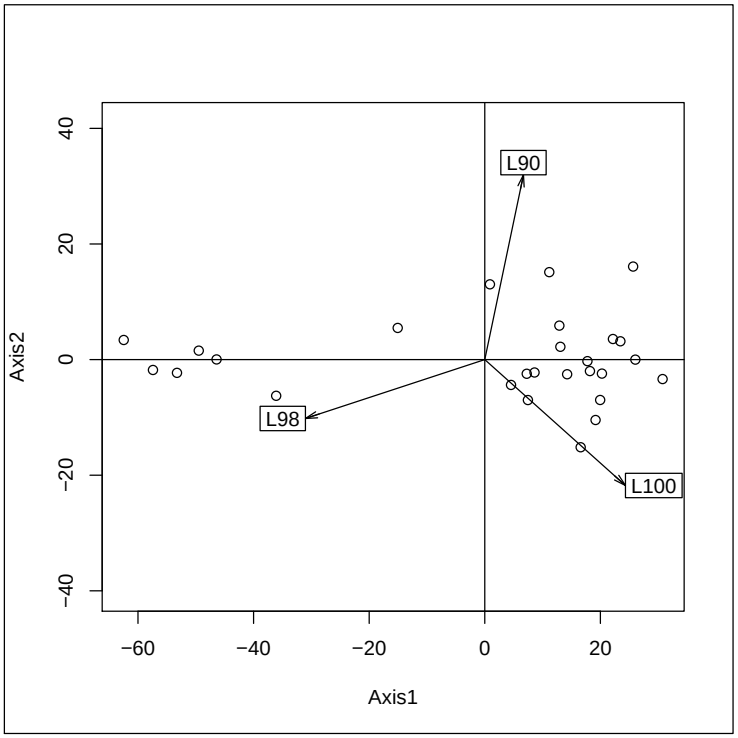
Polymorphisme enzymatique du locus PGM-2* dans 27 échantillons de poissons *Leuciscus cephalus*. Chaque échantillon compte en moyenne 30 individus soit 60 gènes. Le locus compte 3 allèles et le tableau donne les fréquences alléliques dans chaque échantillon. Voir tdr71.

Source : Guinand, B., Y. Bouvet, and B. Brohon. 1996. Spatial aspects of genetic differentiation of the European chub in the Rhone River basin. *Journal of Fish Biology*, **49** :714-726.

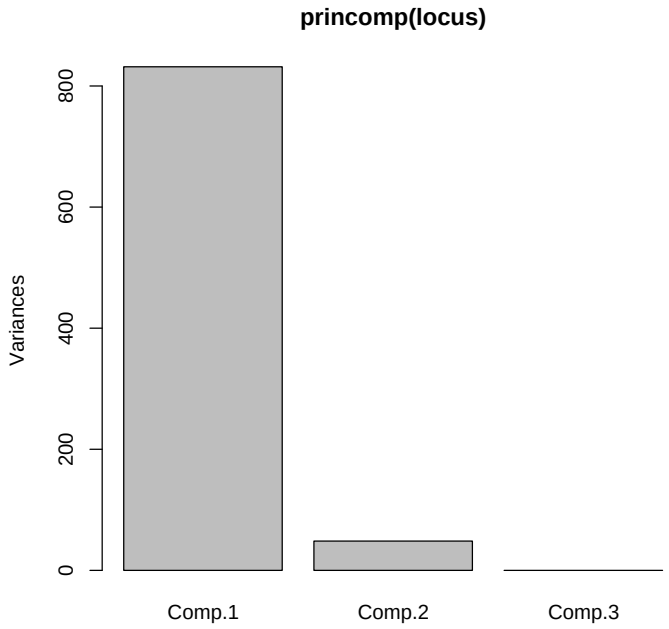
```
triangle.plot(locus, addax = T)
```



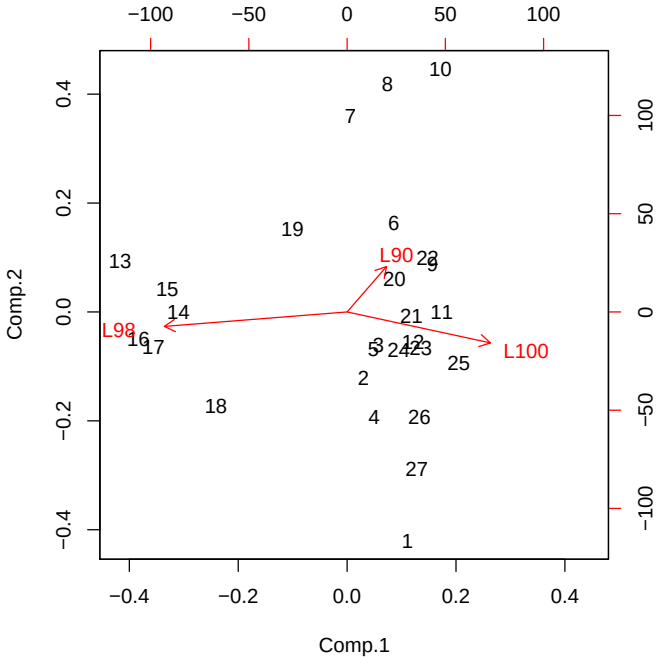

```
pca2 <- dudi.pca(locus, scal = FALSE, scann = FALSE, nf = 2)
plot(pca2$li, asp = 1)
abline(h = 0)
abline(v = 0)
s.arrow(40 * pca2$c1, add.p = TRUE)
```




```
plot(princomp(locus))
```




```
biplot(princomp(locus))
```



À discuter.

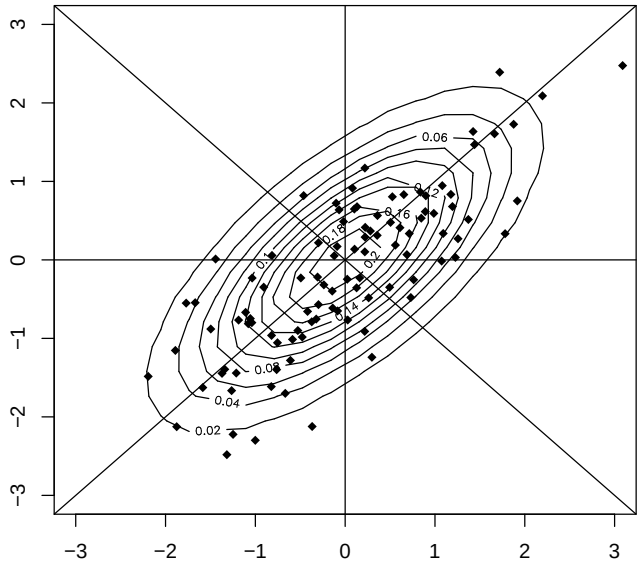
3 Les axes principaux d'une loi normale

Représenter un échantillon et la densité de probabilité binormale par courbes de niveaux :

```
library(mvtnorm)
echa <- mvrnorm(100, mu = c(0, 0), Sigma = matrix(c(1, 0.7, 0.7,
    1), 2, 2))
echa <- as.data.frame(echa)
names(echa) <- c("x", "y")
xg <- seq(-3, 3, le = 20)
yg <- seq(-3, 3, le = 20)
xyg <- expand.grid(x = xg, y = yg)
xyg[1:5, ]
```

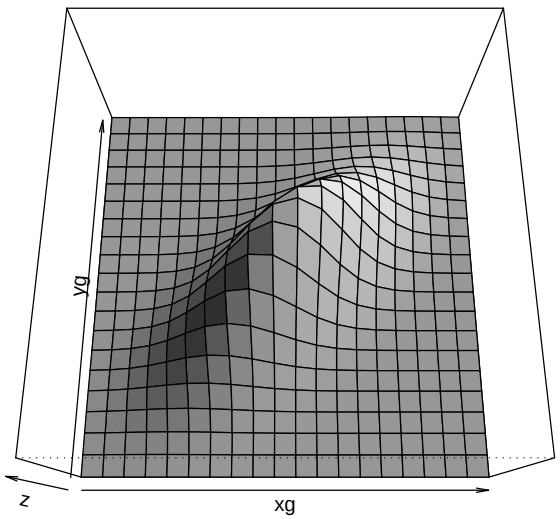
```
          x  y
1 -3.000000 -3
2 -2.684211 -3
3 -2.368421 -3
4 -2.052632 -3
5 -1.736842 -3
```

```
z <- dmvnorm(xyg, me = c(0, 0), sigma = matrix(c(1, 0.7, 0.7, 1),
    2, 2))
z <- matrix(z, nrow = 20)
contour(xg, yg, z)
points(echa$x, echa$y, pch = 18)
abline(v = 0)
abline(h = 0)
abline(0, 1)
abline(0, -1)
```



Représenter la densité de probabilité binormale en perspective :

```
persp(xg, yg, z, theta = 0, phi = 70, expand = 0.5, ltheta = 120,
    shade = 0.75)
```



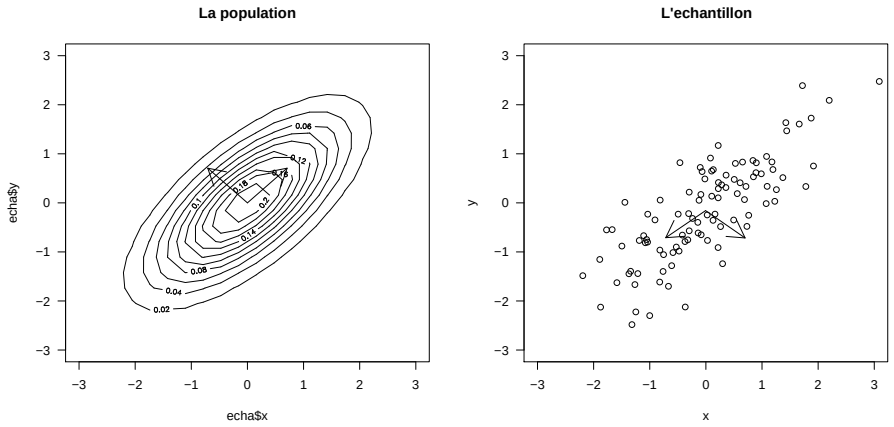
On a créé un échantillon aléatoire simple d'une loi normale multivariée de moyenne (0,0) et de matrice de variances-covariances :

$$\mathbf{C} = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix}$$

La densité s'écrit :

$$g(x,y) = \frac{1}{\sqrt{2\pi^2 \det(\mathbf{C})}} \mathrm{e}^{-\frac{1}{2}\begin{bmatrix} x & y \end{bmatrix} \mathbf{C}^{-1} \begin{bmatrix} x \\ y \end{bmatrix}}$$

```
old.par <- par(no.readonly = TRUE)
par(mfrow = c(1, 2))
plot(echa$x, echa$y, xlim = c(-3, 3), ylim = c(-3, 3), type = "n",
    las = 1, main = "La population")
contour(xg, yg, z, add = TRUE)
arrows(0, 0, 1/sqrt(2), 1/sqrt(2))
arrows(0, 0, -1/sqrt(2), 1/sqrt(2))
pca0 <- prcomp(echa)
plot(echa, xlim = c(-3, 3), ylim = c(-3, 3), las = 1, main = "L'echantillon")
moy <- apply(echa, 2, mean)
arrows(moy[1], moy[2], moy[1] + pca0$rotation[1, 1], moy[1] + pca0$rotation[2,
    1])
arrows(moy[1], moy[2], moy[1] + pca0$rotation[1, 2], moy[1] + pca0$rotation[2,
    2])
par(old.par)
```



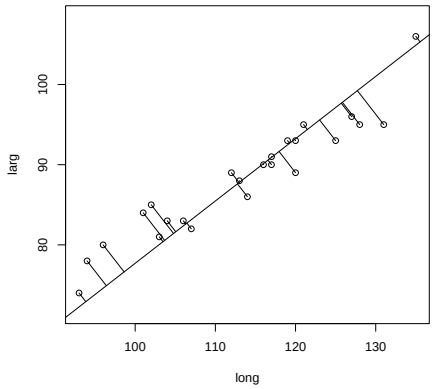
4 Variables non centrées

Reprendre le tableau tortues et extraire longueur et largeur des mâles :

```
tortues <- read.table("http://pbil.univ-lyon1.fr/R/donnees/tortues.txt",
    header = TRUE)
w <- tortues[tortues$sexe == "M", c("long", "larg")]
```

Représenter l'axe principal du nuage non centré et son principe :

```
plot(w, asp = 1)
u <- eigen(as.matrix(t(w) %*% as.matrix(w)), sym = TRUE)$vectors[,
    1]
p <- u[2]/u[1]
abline(0, p)
segments(w[, 1], w[, 2], as.matrix(w) %*% u * u[1], as.matrix(w) %*%
    u * u[2])
```



5 Mode Q et R

On considère un tableau comportant 10 lignes et 3 colonnes consignant l'avis fourni par 10 consommateurs sur la qualité de 3 objets à l'aide du code -1 (avis défavorable) 0 (sans opinion) et +1 (avis favorable). Les 3 objets sont notés A, B, C et les 10 personnes interrogées sont numérotées de 1 à 10.

```
rq <- matrix(c(-1, -1, -1, -1, 0, 1, -1, 0, 0, 1, 1, 1, 0, 0, 1,
    -1, -1, 0, 0, 0, 0, 0, 1, 1, -1, 0, 1, -1, 1, 1), 10, 3, byrow = TRUE)
rq <- as.data.frame(rq)
names(rq) <- LETTERS[1:3]
rq
```

```
   A  B  C
1  -1 -1 -1
2  -1  0  1
3  -1  0  0
4   1  1  1
5   0  0  1
6  -1 -1  0
7   0  0  0
8   0  1  1
9  -1  0  1
10 -1  1  1
```

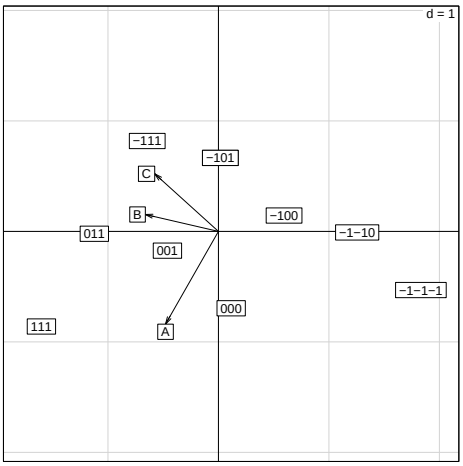
Représenter la projection du nuage de 10 points de $\mathbb{R}^3$ *centré par colonnes* sur ses deux premiers axes principaux et la projection de la base canonique sur ce même plan :

```
a <- apply(rq, 2, mean)
rq1 <- sweep(rq, 2, a)
rq1
```

```
     A    B    C
1  -0.5 -1.1 -1.5
2  -0.5 -0.1  0.5
3  -0.5 -0.1 -0.5
4   1.5  0.9  0.5
5   0.5 -0.1  0.5
6  -0.5 -1.1 -0.5
```

```
7   0.5 -0.1 -0.5
8   0.5  0.9  0.5
9  -0.5 -0.1  0.5
10 -0.5  0.9  0.5
```

```
as.character.data.frame <- function(df) {
    f <- function(x) unlist(paste(x, sep = "", collapse = ""))
    apply(df, 1, f)
}
s.label(dudi.pca(rq1, cent = F, scal = F, scan = F)$li, lab = as.character.data.frame(rq))
s.arrow(dudi.pca(rq1, cent = F, scal = F, scan = F)$c1, add.plot = T)
```

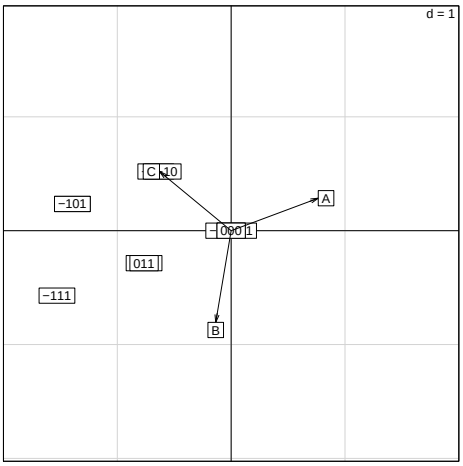


Représenter la projection du nuage de 10 points de $\mathbb{R}^3$ *centré par lignes* sur ses deux premiers axes principaux et la projection de la base canonique sur ce même plan :

```
b <- apply(rq, 1, mean)
rq2 <- sweep(rq, 1, b)
rq2
```

```
           A          B         C
1   0.0000000  0.0000000 0.0000000
2  -1.0000000  0.0000000 1.0000000
3  -0.6666667  0.3333333 0.3333333
4   0.0000000  0.0000000 0.0000000
5  -0.3333333 -0.3333333 0.6666667
6  -0.3333333 -0.3333333 0.6666667
7   0.0000000  0.0000000 0.0000000
8  -0.6666667  0.3333333 0.3333333
9  -1.0000000  0.0000000 1.0000000
10 -1.3333333  0.6666667 0.6666667
```

```
s.label(dudi.pca(rq2, cent = F, scal = F, scan = F)$li, lab = as.character.data.frame(rq),
    xlim = c(-2, 2))
s.arrow(dudi.pca(rq2, cent = F, scal = F, scan = F)$c1, add.plot = T)
```



Détailler les calculs et donner un sens à ces deux figures.
Expliquer ce résultat :

```
eigen(as.matrix(rq2) %*% t(as.matrix(rq2))/2)$values
```

```
[1]  4.097168e+00  5.694991e-01  9.033683e-17  3.800998e-17  8.426918e-19
[6]  0.000000e+00 -8.223493e-33 -2.446522e-32 -1.465512e-16 -1.944285e-16
```

```
eigen(t(as.matrix(rq2)) %*% as.matrix(rq2)/2)$values
```

```
[1]  4.097168e+00  5.694991e-01 -8.090263e-17
```

```
eigen(cov(t(rq)))$values
```

```
[1]  4.097168e+00  5.694991e-01  1.771070e-16  1.517250e-17  2.435008e-19
[6]  0.000000e+00 -2.054311e-32 -2.854187e-32 -1.584636e-16 -3.669611e-16
```

```
eigen(as.matrix(rq1) %*% t(as.matrix(rq1))/9)$values
```

```
[1]  1.084503e+00  3.458286e-01  1.141131e-01  2.178898e-16  5.738041e-17
[6]  3.030906e-17  7.048226e-18  3.102043e-18 -5.382376e-18 -8.633484e-18
```

```
eigen(t(as.matrix(rq1)) %*% as.matrix(rq1)/9)$values
```

```
[1] 1.0845028 0.3458286 0.1141131
```

```
eigen(cov(rq))$values
```

```
[1] 1.0845028 0.3458286 0.1141131
```

6 Valeurs propres doubles

Voir[1].

6.1 Imprécision des calculs numériques

Diagonaliser les matrices de type :

$$\mathbf{R} = \begin{bmatrix} 1 & a & a \\ a & 1 & a \\ a & a & 1 \end{bmatrix}$$

```
a <- 0.8
R <- matrix(c(1, a, a, a, 1, a, a, a, 1), 3, 3)
R
```

[1] Cehessat, R. (1976) Exercices commentés de Statistique et Informatique Appliquée. Dunod, 418 p.

```
     [,1] [,2] [,3]
[1,]  1.0  0.8  0.8
[2,]  0.8  1.0  0.8
[3,]  0.8  0.8  1.0
```

```
b <- eigen(R, sym = TRUE)$values
b[2] == b[3]
```

```
[1] FALSE
```

```
b
```

```
[1] 2.6 0.2 0.2
```

```
print(b, dig = 20)
```

```
[1] 2.6000000000000001 0.2000000000000011 0.2000000000000001
```

En informatique les valeurs propres multiples n'existent pas !

6.2 Certificat d'études

	1	2	3	4	5	6	7	8	9	10	11
Maths	1	2	3	4	5	6	7	8	9	10	11
Musique	6	1	4	5	3	2	9	7	8	10	11
Sanscrit	2	6	5	3	4	1	8	9	7	10	11

TAB. 1 – Classement au certificat d'études

	1	2	3	4	5	6	7	8	9	10	11
Natation	3	4	5	6	7	8	9	10	11	1	2

TAB. 2 – Résultats pour la natation

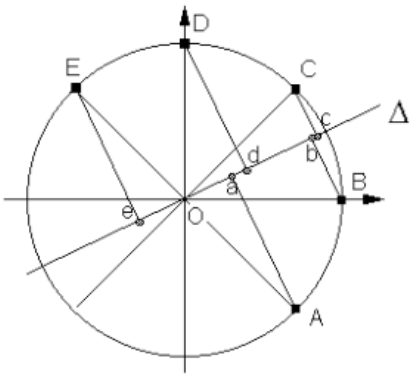
1. On dispose (table 1) du classement de 11 individus pour trois matières du certificat d'études (nouveau programme). Effectuer l'ACP normée du tableau 11 individus - 3 variables. La base des axes principaux est-elle unique ? Tracer le graphe canonique.
2. Un auditeur libre a eu des notes qui lui auraient donné les rangs 8 en maths, 2 en musique et 1 en sanscrit. Le situer par rapport à l'ensemble des individus.
3. Le lendemain, on dispose des classements des mêmes candidats en natation (table 2). Positionner le point natation sur la première composante principale. Aurait-on obtenu le même résultat en analysant un tableau 11 individus - 4 variables ?

7 Exercices

1. $A = (1, 0)$, $B = (\frac{\sqrt{3}}{2}, \frac{1}{2})$, $C = (0, 1)$ sont trois points du plan et a, b et c sont leur projection sur un vecteur unitaire du plan (métrique canonique). Trouver le vecteur qui maximise :

$$Oa^2 + Ob^2 + Oc^2$$

2. On considère 5 points $A, \cdots, E$ sur le cercle unité définis par les angles polaires $\frac{-\pi}{4}$, 0, $\frac{\pi}{4}$, $\frac{\pi}{2}$ et $\frac{3\pi}{4}$. Soit Δ une droite passant par l'origine et $a, \cdots, e$ les projections orthogonales des 5 points initiaux sur Δ. Quelle est la droite Δ qui minimise la quantité $f(\Delta) = Oa^2 + Ob^2 + Oc^2 + Od^2 + Oe^2$? Quel est le maximum atteint ? Quelle est la droite Δ qui minimise la quantité $f(\Delta)$? Quel est le minimum atteint ?



3. Calculer et dessiner le premier axe principal du nuage de 8 points de $\mathbb{R}^2$ défini par :

$$\begin{bmatrix} 2 & 5 & 8 & 3 & -2 & -5 & -8 & -3 \\ -4 & 0 & 4 & 4 & 4 & 0 & -4 & -4 \end{bmatrix}$$

4. On considère les 9 points de $\mathbb{R}^2$ de coordonnées :

$$\begin{bmatrix} M_1 & M_2 & M_3 & M_4 & M_5 & M_6 & M_7 & M_8 & M_9 \\ -4 & -3 & -2 & -1 & 0 & 1 & 2 & 3 & 4 \\ -3 & -4 & -1 & -2 & 0 & 4 & 3 & 2 & 1 \end{bmatrix}$$

Soit O l'origine du plan, $\mathbf{u}$ un vecteur unitaire, m_i la projection orthogonale de M_i sur $\mathbf{u}$, $h(\mathbf{u})$ la somme des carrés des longueurs de Om_i. Le vecteur qui maximise $h(\mathbf{u})$ est-il porté par la première bissectrice ? La valeur maximale atteinte est-elle 118 ?

5. Soient les deux variable à 6 valeurs $x = (0, 1, 0, 1, 1, 0)$ et $y = (0, 1, 1, 0, 1, 0)$. Calculer moyennes, variances et covariances. Dessiner le nuage centré et placer les deux droites de régression et les deux axes principaux.
6. On considère les 10 points M_i de coordonnées :

$$\begin{bmatrix} i & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ x_i & -2 & -1 & 0 & 1 & 2 & -2 & -1 & 0 & 1 & 2 \\ y_i & 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 \end{bmatrix}$$

Caculer $m(\mathbf{x})$, $m(\mathbf{y})$, $v(\mathbf{x})$, $v(\mathbf{y})$ et $c(\mathbf{x}, \mathbf{y})$. Calculer les deux droites de régression et l'axe principal. Tracer le nuage et ses droites caractéristiques.

7. On considère les 10 points P_i de coordonnées :

$$z_i = x_i \cos\alpha + y_i \sin\alpha$$

$$t_i = x_i \sin\alpha - y_i \cos\alpha$$

Caculer $m(\mathbf{z})$, $m(\mathbf{t})$, $v(\mathbf{z})$, $v(\mathbf{t})$ et $c(\mathbf{z}, \mathbf{t})$ pour α quelconque. Tracer les nouveaux nuages et ses trois droites caractéristiques pour $\alpha = 60\circ$. Montrer que, en général, l'axe principal est conservé par rotation du nuage.