Fiche TD avec le logiciel R : tdr60

Trois variables pour commencer

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

Des premiers exercices pour voir que l'analyse des données réunit *des données numériques*, donc une réalité particulière, *des supports théoriques*, essentiellement de la géométrie euclidienne et *des procédures* qui permettent une interaction entre les deux.

Table des matières

1 Un locus et trois allèles

Utiliser la liste chevaine du paquet ade4 présentée dans la fiche :

`http://pbil.univ-lyon1.fr/R/pps/pps054.pdf`

```
library(ade4)
data(chevaine)
x <- chevaine$tab[, 1:3]
head(x)
```

```
    PGM-2*.090 PGM-2*.098 PGM-2*.100
P01      0.017      0.150      0.833
P02      0.083      0.216      0.701
P03      0.107      0.179      0.714
P04      0.067      0.200      0.733
P05      0.103      0.190      0.707
P06      0.179      0.125      0.696
```

```
names(x) <- c("A90", "A98", "A100")
n <- chevaine$eff
```

On ne s'intéresse qu'aux trois premières variables du tableau.

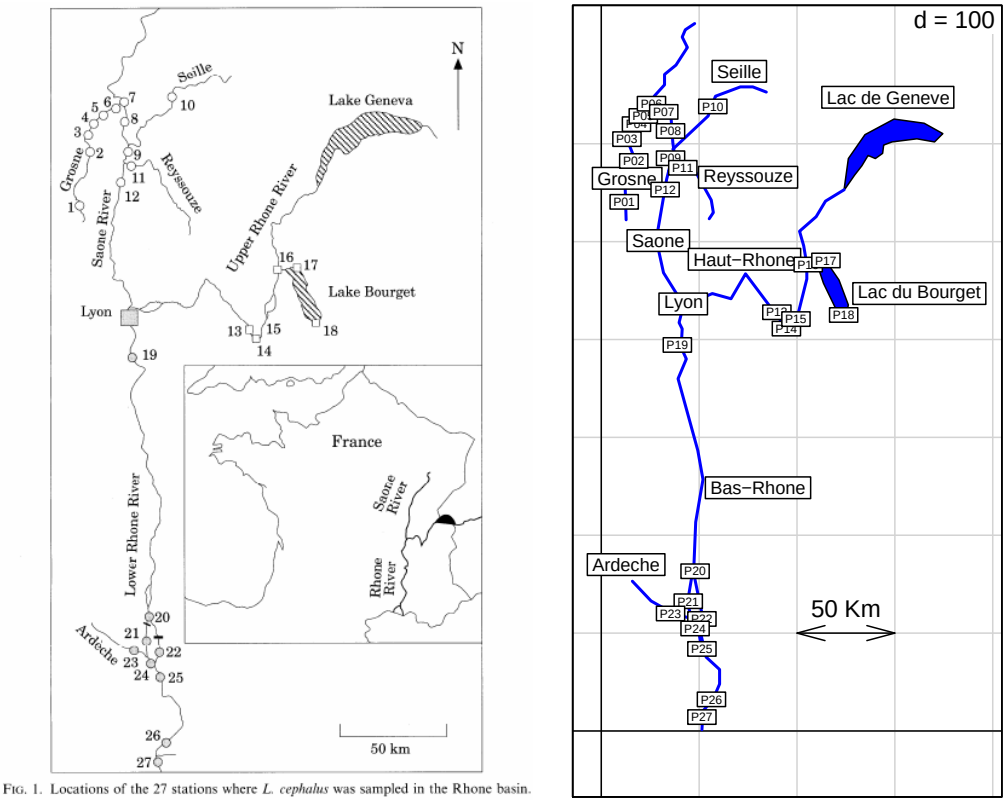


FIG. 1. Locations of the 27 stations where *L. cephalus* was sampled in the Rhone basin.

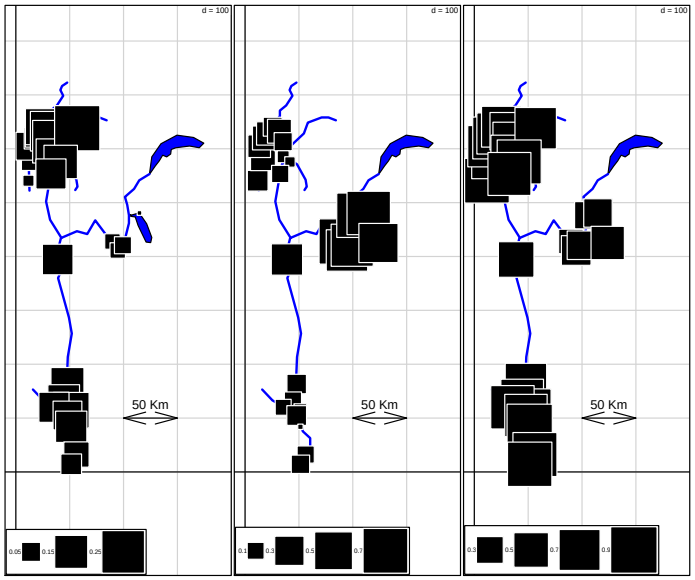
FIG. 1 – Espace concret : 25 stations.

Les lignes sont des stations où ont été capturés des chevaines et les colonnes donnent pour chaque échantillon les fréquences alléliques des allèles 90 98 et 100 pour le locus `PGM-2*`. Les effectifs de l'échantillon sont dans le vecteur `n`. Tous les détails sont dans l'article [2]. Ces effectifs concernent-ils les individus ou les gênes (diploïdie) ou ni l'un ni l'autre ?

Consulter la carte de documentation de l'objet pour récupérer une fonction qui dessine un fond de carte. La disposition des stations dans le bassin rhôdanien est décrite dans l'article (à gauche) et schématisée par la fonction (à droite). Refaire ce schéma.

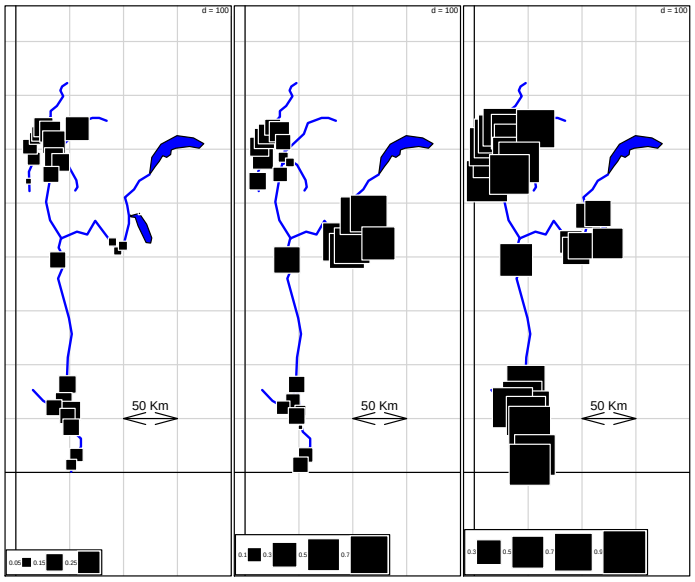
Cartographier les données :

```
xy <- chevaine$coo$sta
par(mfrow = c(1, 3))
for (k in 1:3) {
    fun.chevaine(F)
    s.value(xy, x[, k], add.p = T, csi = 3)
}
```



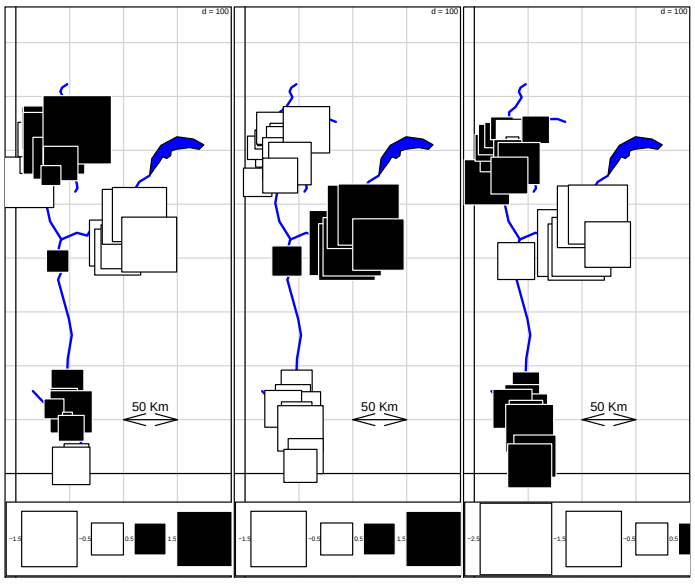
On change un paramètre. Quelle différence avec le précédent ? La représentation graphique des données est une partie essentielle de leur analyse. On ne prend jamais assez de précautions à ce niveau. Un dessin contient toujours une intention, maîtrisée ou non :

```
xy <- chevaine$coo$sta
par(mfrow = c(1, 3))
for (k in 1:3) {
    fun.chevaine(F)
    s.value(xy, x[, k], add.p = T, csi = 3, zmax = 1)
}
```



La cartographie par valeur est un procédé graphique pauvre : l'œil apprécie mal les surfaces et ceci invalide les camemberts. Le centrage préalable améliore la situation.

```
x0 <- scale(x)
par(mfrow = c(1, 3))
for (k in 1:3) {
    fun.chevaine(F)
    s.value(xy, x0[, k], add.p = T, zmax = 1, csi = 3)
}
```

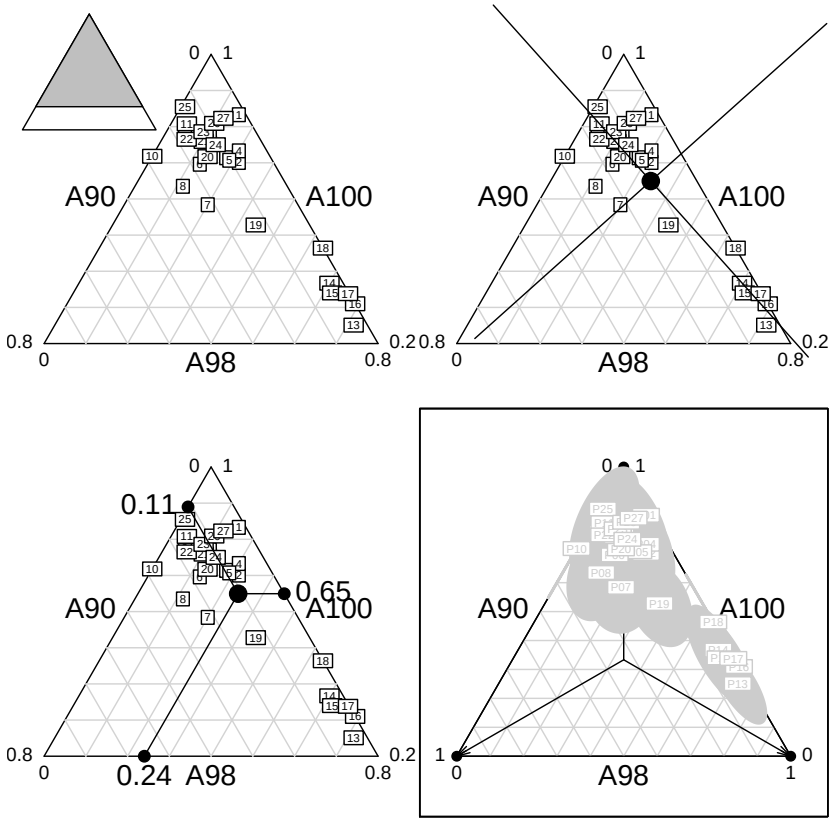


Quand on connaît deux des trois valeurs la troisième est imposée ($x+y+z=1$). On dit que les données sont de dimensions 2. L'espace concret est aussi de dimension 2. On vient de représenter les données dons l'espace concret. On peut

aussi représenter les stations dans l'espace abstrait, celui des données. Pour faire le tour de la représentation triangulaire, sorte de porte d'entrée du système de la statistique euclidienne, voir :

http://pbil.univ-lyon1.fr/R/fichestd/ter1.pdf

```
par(mfrow = c(2, 2))
triangle.plot(x, clab = 0.75)
triangle.plot(x, clab = 0.75, adda = T, show = F)
triangle.plot(x, clab = 0.75, addm = T, show = F)
id3 <- as.data.frame(diag(1, 3))
names(id3) <- names(x)
w <- triangle.plot(id3, show.pos = F)
row.names(w) <- names(x)
s.multinom(w, x, add.plot = T, n.sample = 2 * n, axesell = F, clabelcat = 0)
```



Commenter les éléments de cette figure. Synthétiser l'information acquise. Nous avons deux images très différentes d'une même information : la variabilité des profils alléliques structuré dans l'espace fluvial. Dans l'un, on voit la variabilité des fréquences dans la variabilité des coordonnées. Dans l'autre on devine les coordonnées dans la variabilité des fréquences. Il existe une méthode qui superpose les deux systèmes. Les plus curieux feront :

```
plot(procuste(xy, x))
```

Le résultat est surprenant mais le contenu théorique plus sérieux.

2 Chercher une expression graphique de l'information

Analyser des données, c'est d'abord les regarder. Considérer la liste sarcelles.

```
data(sarcelles)
sarcelles$tab
```

```
             Jui Aou Sep Oct Nov Dec Jan Fev Mar Avr Mai Jun
Fr._medit.     0  12  44  44  48  69 179 500 147   4   0   0
Fr._conti.     4   4  30  16  20  11  17  75  70   1   0   0
Fr._Atlan.     0   7   9  13   7  14  18  74  16   0   0   0
P._Iberi.      1   0   0   3   5   6  11  98   6   0   0   0
Ital._N.       0   4  38  18  21   9  16  87 218  51   2   0
Ital._S.       0   1   2   5   3   7   1  11  14   2   0   0
Ang.Irl.       1   0   3   2   0   7   5   1   0   1   0   0
Bel.Holl.      0   7  28  17   9   4   9   9  10   2   0   0
Sui._All.S.    0   4   8  12   2   3   5   5   1   0   0   1
Allem.N.       1  12  20  13   9   2   0   0   2   1   0   0
Aut.Tch.H.Po.  1  43  31   7   2   1   1   1   4   5   0   1
Scand.S.       4  68  53  15   3   0   0   0   0   5  25   7
Scan.N.        0  14   3   0   0   0   0   0   0   0   7   1
URSS           3 184 105  34   5   0   0   0   0  11  83  13
```

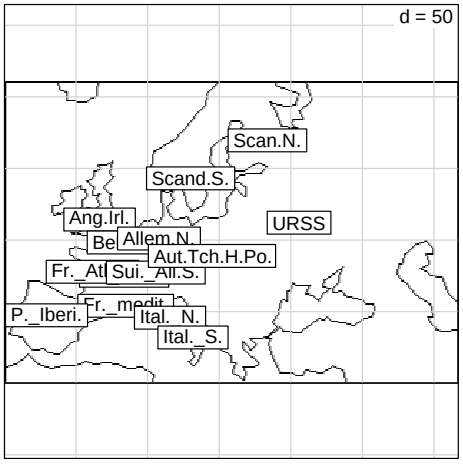
Les données [4] forment une table de contingence croisant régions et mois de capture pour 3049 reprises de bagues de Sarcelles d'hiver. Elles sont publiées dans [7]. Dans le tableau à $I = 14$ lignes et $J = 12$ colonnes sont inscrits les effectifs n_{ij} de Sarcelles d'hiver (*Anas C. Crecca*) dont la bague a été récupérée dans la région i au cours du mois j ($n = 3049$). Ces bagues sont posées en hiver, en Camargue, et retrouvées après la mort de l'oiseau, en général par la chasse.

Récupérer une image grossière de l'Europe :

```
library(pixmap)
sarcelles.pnm <- read.pnm(system.file("pictures/sarcelles.pnm",
    package = "ade4"))
```

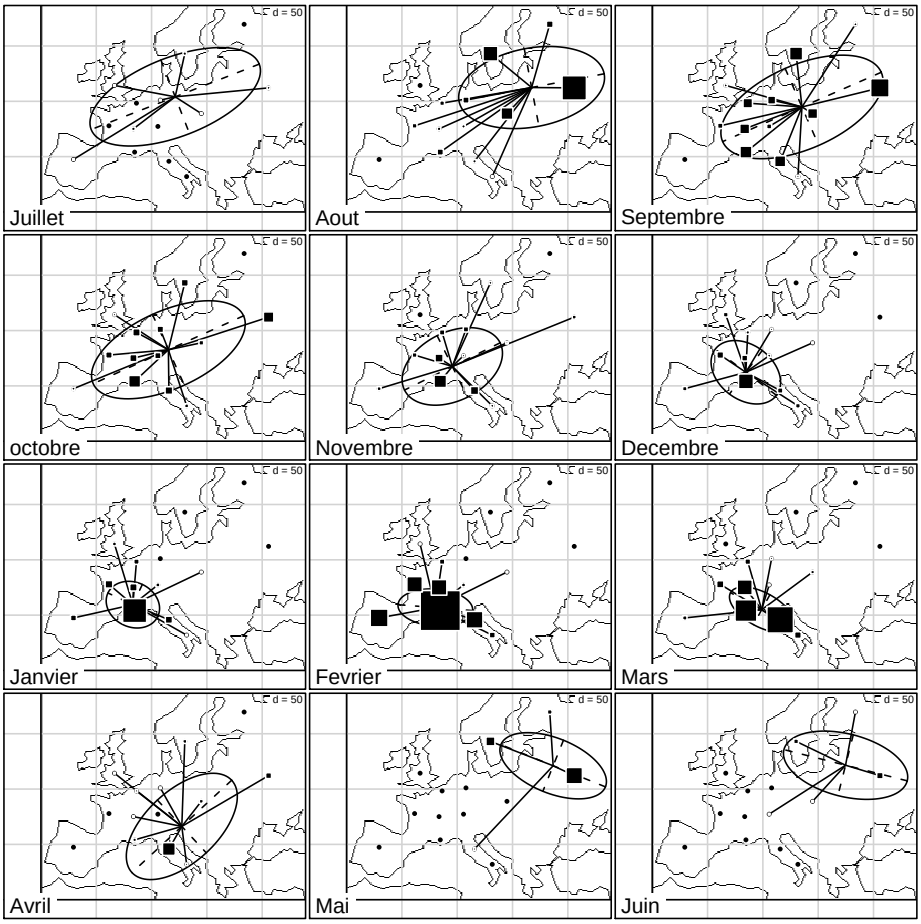
Contrôler les étiquettes :

```
s.label(sarcelles$xy, pixmap = sarcelles.pnm, xlim = c(0, 317),
    ylim = c(0, 212), label = row.names(sarcelles$tab))
```



Représenter les données :

```
par(mfrow = c(4, 3))
for (i in 1:12) {
    s.distri(sarcelles$xy, sarcelles$tab[, i], pixmap = sarcelles.pnm,
        sub = sarcelles$col.names[i], clab = 0, csub = 2)
    s.value(sarcelles$xy, sarcelles$tab[, i], add.plot = TRUE, cleg = 0,
        csi = 2, zmax = 500)
}
```



Commenter. L'ellipse de dispersion, qui a ici beaucoup de signification biologique (la migration d'automne est plus un regroupement qu'un déplacement), est un procédé graphique d'expression de la variabilité à deux dimensions.

Dans chaque figure on a un vecteur $\mathbf{x}$ à $n = 14$ composantes (l'abscisse du centre de la région), un vecteur $\mathbf{y}$ à $n = 14$ composantes (l'ordonnée du centre de la région) et un vecteur $\mathbf{z}$ à $n = 14$ composantes (l'effectif des reprises de bagues dans la région).

On ne tient compte que des fréquences formant le vecteur $\mathbf{f}$ avec $z_\bullet = \sum_{j=1}^{j=n} z_j$ et $f_i = \frac{z_i}{z_\bullet}$. La matrice diagonale :

$$\mathbf{D} = \text{Diag}\,(f_1, f_2, ..., f_s)$$

est un objet conceptuel qui permet de remplacer les signes somme par des écritures matricielles :

Le centre de gravité du nuage des oiseaux a pour coordonnées les moyennes :

$$\mathrm{m}(\mathbf{x}) = \overline{\mathbf{x}} = \sum_{j=1}^{j=n} f_j x_j = \mathbf{x}^T\mathbf{f} = \mathbf{x}^T\mathbf{D}\mathbf{1}_n$$
$$\mathrm{m}(\mathbf{y}) = \overline{\mathbf{y}} = \sum_{j=1}^{j=n} f_j y_j = \mathbf{y}^T\mathbf{f} = \mathbf{y}^T\mathbf{D}\mathbf{1}_n$$

On utilise alors les variables centrées :

$$\mathbf{X} = \mathbf{x} - \mathrm{m}(\mathbf{x})\mathbf{1}_n$$
$$\mathbf{Y} = \mathbf{y} - \mathrm{m}(\mathbf{y})\mathbf{1}_n$$

Sur chaque axe, la dispersion des oiseaux est mesurée par la variance :

$$\mathrm{v}(\mathbf{x}) = \sum_{j=1}^{j=n} f_j (x_j - \mathrm{m}(\mathbf{x}))^2 = \mathbf{X}^T\mathbf{D}\mathbf{X}$$
$$\mathrm{v}(\mathbf{y}) = \sum_{j=1}^{j=n} f_j (y_j - \mathrm{m}(\mathbf{y}))^2 = \mathbf{Y}^T\mathbf{D}\mathbf{Y}$$

Mais cette dispersion n'est pas quelconque. Le nuage peut être allongé, ce qui renvoie à la covariance :

$$\mathrm{cov}(\mathbf{x},\mathbf{y}) = \sum_{j=1}^{j=n} f_j (x_j - \mathrm{m}(\mathbf{x}))(y_j - \mathrm{m}(\mathbf{y})) = \mathbf{X}^T\mathbf{D}\mathbf{Y}$$

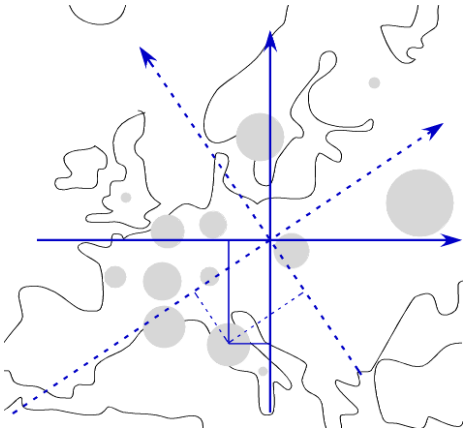
La matrice à 2 lignes et 2 colonnes qui contient les paramètres de dispersion s'écrit très simplement :

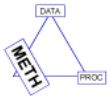
$$\mathbf{C} = \begin{bmatrix} \mathrm{v}(\mathbf{x}) & \mathrm{cov}(\mathbf{x},\mathbf{y}) \\ \mathrm{cov}(\mathbf{y},\mathbf{x}) & \mathrm{v}(\mathbf{y}) \end{bmatrix} = \begin{bmatrix} \mathbf{X},\mathbf{Y} \end{bmatrix}^T \mathbf{D} \begin{bmatrix} \mathbf{X},\mathbf{Y} \end{bmatrix} = \mathbf{Z}^T\mathbf{D}\mathbf{Z}$$

Le raisonnement qui suit est alors fondamental. C'est, pratiquement, celui de K. Pearson[8], dans un article mythique :

`http://pbil.univ-lyon1.fr/R/html/liens/pearson1901.pdf`

On a mesuré la dispersion en longitude et en latitude avec une certaine covariance. On peut la mesurer dans n'importe quelle direction :





On se donne pour caractériser un axe quelconque son vecteur directeur

$$\mathbf{u} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$$

, c'est à dire deux nombres α et β tels que $\alpha^2 + \beta^2 = 1$. Les coordonnées sur ce nouvel axe sont dans le vecteur $\mathbf{Zu}$ qui est centré et a pour variance

$$\mathrm{Q}(\mathbf{u}) = \mathbf{u}^T\mathbf{D}\mathbf{Z}\mathbf{u} = \mathbf{u}^T\mathbf{C}\mathbf{u}$$

La matrice $\mathbf{C}$ est symétrique. Elle a toujours deux valeurs propres $\lambda_1 \geq \lambda_2 \geq 0$ positives ou nulles et deux vecteurs propres orthonormés $\mathbf{u}_1$ et $\mathbf{u}_2$. La variabilité sur le nouvel axe avec ses coordonnées dans l'ancien système s'écrit :

$$\mathrm{Q}(\mathbf{u}) = \alpha^2\mathrm{v}(\mathbf{x}) + 2\alpha\beta\mathrm{cov}(\mathbf{x},\mathbf{y}) + \beta^2\mathrm{v}(\mathbf{y})$$

Si on utilise ses coordonnées a et b dans la nouvelle base des vecteurs propres, cette quantité devient :

$$\mathrm{Q}(\mathbf{u}) = \mathbf{u}^T\mathbf{C}\mathbf{u} = \mathbf{u}^T\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T\mathbf{u} = \lambda_1 a^2 + \lambda_2 b^2$$

Le premier vecteur propre est celui qui maximise la quantité $\mathrm{Q}(\mathbf{u})$ (le maximum est λ_1), le second est celui qui minimise cette quantité (le minimum est λ_2). Ces deux vecteurs sont appelés les axes principaux du nuage. L'ellipse de dispersion est celle qui est définie par le centre de gravité et les deux axes $k\sqrt{\lambda_1}\mathbf{u}_1$ et $k\sqrt{\lambda_2}\mathbf{u}_2$. k, un coefficient de taille est libre et vaut par défaut 1.5. Des détails sont dans :

`http://pbil.univ-lyon1.fr/R/querep/qr3.pdf`

Faut-il savoir ce qu'est un vecteur propre pour utiliser des ellipses d'inertie ? Le débat est loin d'être clos !

3 Longueur, largeur et hauteur des carapaces de tortues

©Michel Pilon 2005

`http://parcours.pilonm.org/pictures/reptiles/tortuePeinte3.jpg`

48 tortues peintes (*Chrysemys picta marginata*) ont été capturées par J.E. Mosimann, le même jour dans un même étang, et appartenaient donc à la même population. Les carapaces ont été mesurées et publiées par P. Jolicœur et J.E. Mosimann [6]. On trouve ces données dans un paquet de R :

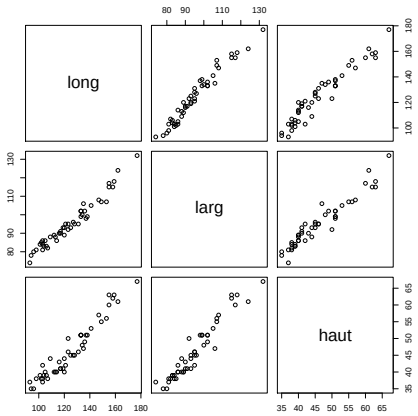
```
library(Flury)
data(turtles)
```

A défaut, utiliser le fichier :

```
tortues <- read.table(url("http://pbil.univ-lyon1.fr/R/donnees/tortues.txt"),
    h = T)
```

Séparer les mesures quantitatives (en mm) et la variable qualitative qui donne le sexe de l'animal :

```
xyz <- tortues[, -4]
sexe <- tortues[, 4]
pairs(xyz)
```



Longueur, largeur et hauteur sont fortement corrélées.

La question sous-jacente à l'origine des données est *Isométrie ou allométrie ?* L'allométrie est un changement dans les proportions du corps d'un animal au cours de sa croissance, par suite du développement plus rapide ou plus lent de l'une de ses parties, voir :

`http://www.unice.fr/LEML/coursJDV/morpho/morpho6-1.htm`

La carapace change-t-elle de forme pendant la croissance ? Dans le cas contraire, à une constante multiplicative près, toutes les carapaces sont identiques à une forme canonique (une tortue *canon*, en quelque sorte !)

Peut-on trouver une tortue de synthèse (α, β, γ) reproduite approximativement par la tortue i sous la forme ($\mathbf{X}$ est le tableau de données) :

$$(\mathbf{X}[i,1], \mathbf{X}[i,1], \mathbf{X}[i,1]) \approx u_i(\alpha, \beta, \gamma)$$

Si on cherche une approximation, il faut se donner un critère d'erreur, sous la forme d'une norme dans l'espace des tableaux. On passe facilement de l'espace des vecteurs à l'espace des matrices en vectorisant les matrices, ce qui est explicite dans R :

```
a <- matrix(1:12, 4, )
a
     [,1] [,2] [,3]
[1,]    1    5    9
[2,]    2    6   10
[3,]    3    7   11
[4,]    4    8   12
as.numeric(a)
 [1]  1  2  3  4  5  6  7  8  9 10 11 12
```

A deux matrices à n lignes et p colonnes $\mathbf{A} = [a_{ij}]$ et $\mathbf{B} = [b_{ij}]$, on associe le nombre :

$$\mathbf{A} \bullet \mathbf{B} = \sum_{i=1}^{i=n} \sum_{j=1}^{i=p} a_{ij} b_{ij} = \text{trace}\left(\mathbf{A}^T \mathbf{B}\right)$$

On obtient un produit scalaire (le produit scalaire canonique des matrices réelles à n lignes et p colonnes).

On a donc une norme qui mesure l'écart entre deux matrices :

$$d^2\left(\mathbf{A}, \mathbf{B}\right) = \|\mathbf{B} - \mathbf{A}\|^2 = (\mathbf{B} - \mathbf{A}) \bullet (\mathbf{B} - \mathbf{A}) = \sum_{i=1}^{i=n} \sum_{j=1}^{i=p} (b_{ij} - a_{ij})^2$$

On a donc un tableau de donnée $\mathbf{X}$ dont on cherche une approximation sous la forme :

$$x_{ij} \approx u_i v_j \Leftrightarrow \mathbf{X} \approx \mathbf{u}\mathbf{v}^T$$

Le problème est encore mal posé car, pour une solution éventuelle, on en a une infinité d'autres identiques, en divisant $\mathbf{u}$ par une constante et en multipliant $\mathbf{v}$ par la même constante. On fixe alors des contraintes en optant pour des vecteurs unitaires et en rétablissant une constante d'échelle.

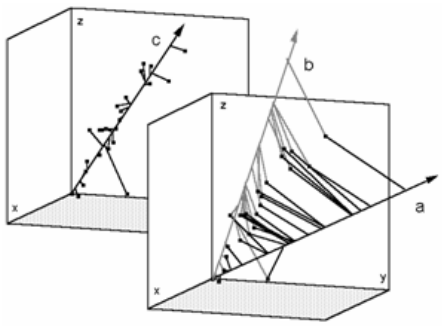
La question devient : trouver $\mathbf{u}$, $\mathbf{v}$ et σ qui vérifie :

$$\|\mathbf{u}\|^2 = \sum_{i=1}^{i=n} u_i^2 = 1 \qquad \|\mathbf{v}\|^2 = \sum_{j=1}^{j=p} v_j^2 = 1$$

et qui rende minimale la quantité :

$$\mathrm{E}(\mathbf{u}, \mathbf{v}, \sigma) = \|\mathbf{X} - \sigma \mathbf{u}\mathbf{v}^T\|^2$$

La recherche de $\mathbf{v}$ correspond à la figure :



La solution unique est donnée par la décomposition en valeurs singulières (***s**ingular **v**alue **d**ecomposition*) de la matrice $\mathbf{X}$. La fonction `svd` assure l'opération. On démontre que, non seulement le problème précédent a une solution unique, mais que la matrice $\mathbf{X}$ se décompose progressivement en matrice de rang 1, sous la contrainte supplémentaire qui impose que les vecteurs $\mathbf{u}_k$ et $\mathbf{v}_k$ soient orthonormés. Ce qu'on écrit, si r est le rang de matrice :

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V} = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^T + \ldots + \sigma_r \mathbf{u}_r \mathbf{v}_r^T$$

Les valeurs singulières de $\mathbf{X}$ sont étroitement liées aux valeurs propres de $\mathbf{X}^T\mathbf{X}$ et $\mathbf{X}\mathbf{X}^T$

```
X <- as.matrix(xyz)
eigen(t(X) %*% X)$values
```

```
[1] 1316320.3629     550.5272     207.1099
```

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

```
 [1]  1.316320e+06  5.505272e+02  2.071099e+02  1.365269e-10  1.294407e-10
 [6]  1.145480e-10  8.241847e-11  5.798237e-11  5.393118e-11  4.930265e-11
[11]  4.292573e-11  4.013752e-11  3.958791e-11  2.847650e-11  2.596052e-11
[16]  1.591352e-11  1.121041e-11  1.076728e-11  6.118401e-12  2.681018e-12
[21]  1.656838e-12  1.066140e-12 -1.280021e-12 -1.590281e-12 -1.919889e-12
[26] -2.194899e-12 -6.181483e-12 -7.962135e-12 -8.842034e-12 -1.018287e-11
[31] -1.274593e-11 -1.281556e-11 -1.295775e-11 -1.439308e-11 -1.547464e-11
[36] -1.571646e-11 -1.661255e-11 -1.880379e-11 -2.699007e-11 -2.947307e-11
[41] -3.124289e-11 -3.496749e-11 -4.053659e-11 -4.413595e-11 -5.808372e-11
[46] -9.704794e-11 -1.012777e-10 -1.024229e-10
```

```
svd(X)$d^2
```

```
[1] 1316320.3629     550.5272     207.1099
```

Démontrer que pour une matrice réelle à n lignes et p colonnes :

$$\text{rang}(\mathbf{X}) = \text{rang}(\mathbf{X}^T) = \text{rang}(\mathbf{X}\mathbf{X}^T) = \text{rang}(\mathbf{X}^T\mathbf{X})$$

Démontrer que, si le rang est r, $\mathbf{X}\mathbf{X}^T$ et $\mathbf{X}^T\mathbf{X}$ ont $r$valeurs propres identiques strictement positives.

Donner les relations qui lient les deux familles de vecteurs propres orthonormées. (On suppose connu le théorème : pour qu'une matrice réelle admette une base orthogonale de vecteurs propres, il faut et il suffit qu'elle soit symétrique). En déduire les relations :

$$\mathbf{X}\mathbf{X}^T = \mathbf{U}\mathbf{\Lambda}_n\mathbf{U}^T \quad \mathbf{\Lambda}_n = diag\left(\lambda_1, ..., \lambda_r, \underbrace{0, ..., 0}_{n-r}\right)$$

$$\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{\Lambda}_p\mathbf{V}^T \quad \mathbf{\Lambda}_p = diag\left(\lambda_1, ..., \lambda_r, \underbrace{0, ..., 0}_{p-r}\right)$$

$$\mathbf{X}\mathbf{X}^T = \mathbf{U}_r\mathbf{\Lambda}\mathbf{U}_r^T \quad \mathbf{X}^T\mathbf{X} = \mathbf{V}_r\mathbf{\Lambda}\mathbf{V}_r^T \quad \mathbf{\Lambda} = diag\left(\lambda_1, ..., \lambda_r\right)$$

$$\mathbf{U}_r = \mathbf{X}\mathbf{V}_r\mathbf{\Lambda}_r^{-1/2} \quad \mathbf{V}_r = \mathbf{X}^T\mathbf{U}_r\mathbf{\Lambda}_r^{-1/2}$$

En déduire que :

$$\mathbf{X} = \mathbf{U}_r\mathbf{D}\mathbf{V}_r^T \quad \mathbf{D} = diag\left(\sqrt{\lambda_1}, ..., \sqrt{\lambda_r}\right)$$

On démontre alors que pour toute matrice $\mathbf{Y}$ à n lignes et p colonnes de rang au plus k ($k \leqslant r$), on a :

$$\|\mathbf{X} - \mathbf{Y}\|^2 = \sum_{i,j}(x_{ij} - y_{ij})^2 \geqslant \lambda_{k+1} + \lambda_{k+2} + ... + \lambda_r$$

Que signifie ce résultat pour $k = 1$?

```
svd(X)$v
```

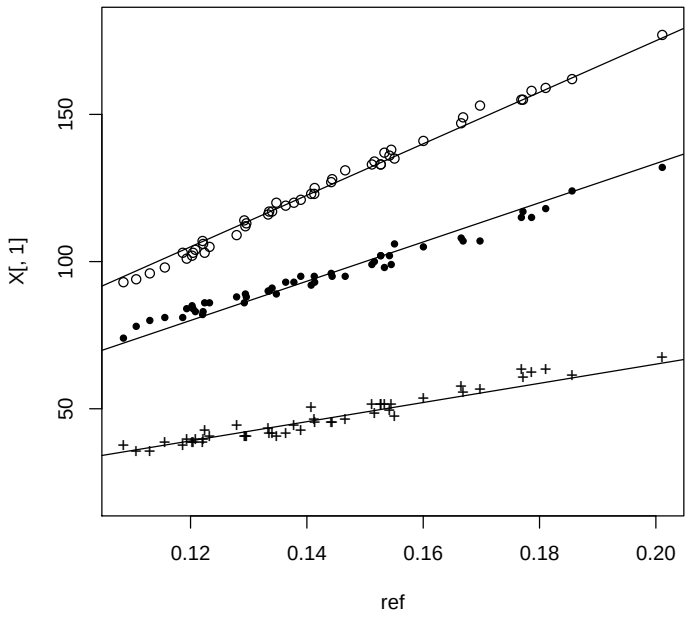
```
           [,1]       [,2]       [,3]
[1,] -0.7627393 -0.5033121 -0.40608574
[2,] -0.5810331  0.8090441  0.08858984
[3,] -0.2839529 -0.3035202  0.90953076
```

```
eigen(t(X) %*% X)$vectors
```

```
           [,1]       [,2]       [,3]
[1,] -0.7627393  0.5033121 -0.40608574
[2,] -0.5810331 -0.8090441  0.08858984
[3,] -0.2839529  0.3035202  0.90953076
```

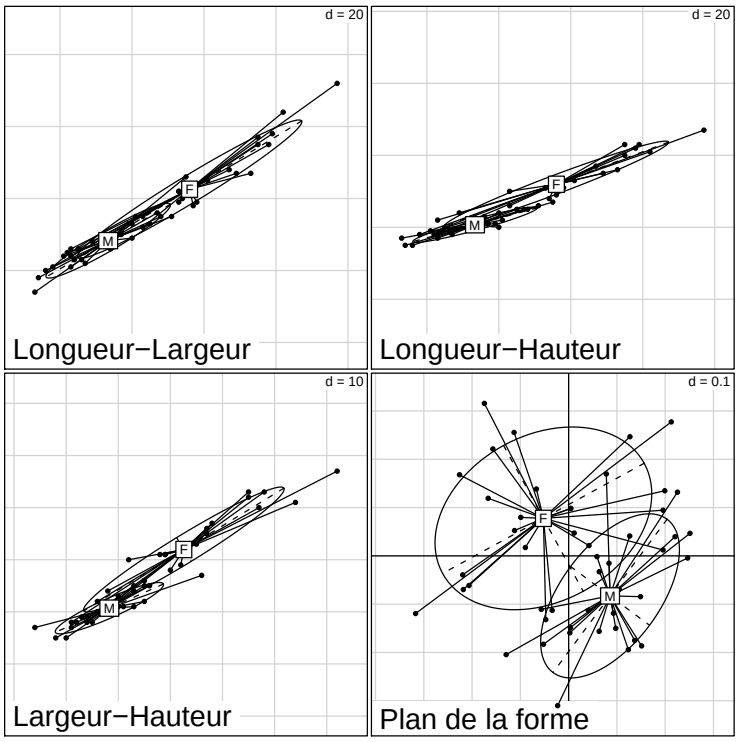
Et vos tortues ? Voilà, voilà !

```
ref <- -svd(X)$u[, 1]
plot(ref, X[, 1], ylim = c(20, 180))
abline(lm(X[, 1] ~ -1 + ref))
points(ref, X[, 2], pch = 20)
abline(lm(X[, 2] ~ -1 + ref))
points(ref, X[, 3], pch = "+")
abline(lm(X[, 3] ~ -1 + ref))
```



Ce n'est satisfaisant qu'en apparence : les résidus sont organisés. Nous avons fait l'impasse sur le dimorphisme sexuel. On peut refaire les modèles par sexe. On peut mettre en évidence les résidus qui sont de dimensions deux :

```
par(mfrow = c(2, 2))
s.class(xyz[, c(1, 2)], sexe, inc = F, sub = "Longueur-Largeur",
    csub = 2)
s.class(xyz[, c(1, 3)], sexe, inc = F, sub = "Longueur-Hauteur",
    csub = 2)
s.class(xyz[, c(2, 3)], sexe, inc = F, sub = "Largeur-Hauteur",
    csub = 2)
s.class(svd(X)$u[, 2:3], sexe, sub = "Plan de la forme", csub = 2)
```

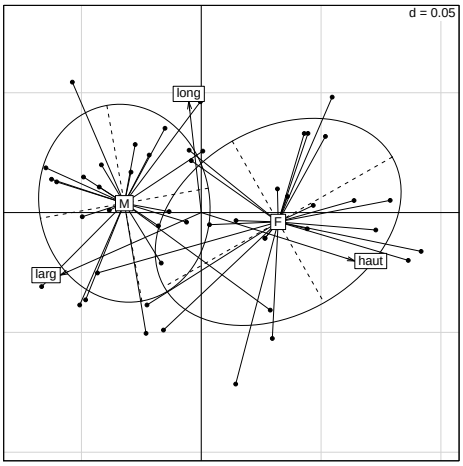


On peut encore travailler en double centrage sur les logarithmes :

```
xyzlog <- log(xyz)
w1 <- t(apply(xyzlog, 1, function(x) x - mean(x)))
svd(w1)$d
```

```
[1] 5.043698e+00 2.469599e-01 2.855608e-15
```

```
w2 <- dudi.pca(as.data.frame(w1), scal = F, scan = F)
s.class(w2$li, sexe)
s.arrow(2 * w2$co, add.plot = T)
```



Expliquer la troisième valeur singulière. Les femelles sont plus grandes, mais à taille égale, elles sont plus hautes et moins larges.

On pourrait encore explorer ce petit jeu de données plein de signification biologique accessible par des pratiques statistiques. P. Jolicœur écrira à la fin de sa carrière un livre très représentatif de cette situation [5].

4 Perspectives

Quelques indications à retenir. Dans un tableau à trois variables, si $x+y+z = 1$, on voit l'intégralité de la variabilité dans la représentation triangulaire. Les données sont dans un plan et on voit ce plan. Si $x+y+z = 0$, pour voir le plan, il faut utiliser la décomposition en valeurs singulières. Le centrage par lignes permet cette opération mais n'a pas toujours un sens.

En morphométrie, le centrage par lignes sur les log met au même endroit les individus de taille différente dans le modèle d'isométrie. Pour les données spatialisées, on peut toujours chercher l'espace concret dans la représentation des données ou utiliser la représentation des données dans l'espace concret. Pour en savoir plus sur la décomposition taille-forme voir l'article de G. Yoccoz [10]. Lui aussi est formé au modèle de la géométrie euclidienne[9].

Le théorème d'approximation des matrices par des matrices de rang plus petit est apparu en statistique avec l'article fondateur de C. Eckart et G. Young (1936)[1].

On trouve sa démonstration p. 556-558 dans l'ouvrage de référence de D.A. Harville, bible du calcul matriciel pour les statisticiens [3].

Pour manipuler d'autres illustrations de la décomposition en valeurs singulières voir :

- les indices boursiers dans `http://pbil.univ-lyon1.fr/R/fichestd/tdr14.pdf`.
- les variables physico-chimiques sur dates-stations dans `http://pbil.univ-lyon1.fr/R/fichestd/tdr332.pdf`

Références

[1] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1 :211–218, 1936.

[2] B. Guinand, Y. Bouvet, and B. Brohon. Spatial aspects of genetic differentiation of the european chub in the rhone river basin. *Journal of Fish Biology*, 49 :714–726, 1996.

[3] D.A. Harville. *Matrix algebra from a statistician's perspective*. Springer, New York, 1997.

[4] L. Hoffmann. Untersuchungen an enten in der camargue. *Ornithologischer Beobatcher*, 57 :35–50, 1960.

[5] P. Jolicœur. *Introduction à la biométrie*. Décarie, Montréal, 3ième edition, 1997.

[6] P. Jolicœur and J.E. Mosimann. Size and shape variation in the painted turtle. a principal component analysis. *Growth*, 24 :339–354, 1960.

[7] J.D. Lebreton. Etude des déplacements saisonniers des sarcelles d'hiver, anas c. crecca l., hivernant en camargue à l'aide de l'analyse factorielle des correspondances. *Compte rendu hebdomadaire des séances de l'Académie des sciences. Paris, D*, III :2417–2420, 1973.

[8] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2 :559–572, 1901.

[9] N. G. Yoccoz. *Le rôle du modèle euclidien d'analyse des données en biologie évolutive*. PhD thesis, Thèse de doctorat, Université Lyon 1, 1988.

[10] N. G. Yoccoz. Morphométrie et analyses multidimensionelles. une revue des méthodes séparant taille et forme. In J.D. Lebreton and B. Asselain, editors, *Biométrie et Environnement*, pages 73–99. Masson, Paris, 1993.