Fiche TD avec le logiciel R : tdr85

Les tortues peintes de JE Mosiman et P Jolicoeur (1960)

A.B. Dufour et I. Amat

Cette fiche a pour objectif de traiter deux questions biologiques via l'algèbre matricielle.

Table des matières

1 Les données biologiques

48 tortues peintes (*Chrysemys picta marginata*) ont été capturées par J.E. Mosiman, le même jour, dans un même étang et appartiennent toutes à la même population. Les carapaces ont été mesurées et publiées par P. Jolicoeur et JE Mosiman [2]. On enregistre les données sous R et on visualise les 6 premières lignes du tableau :

```
tortues <- read.table("https://pbil.univ-lyon1.fr/R/donnees/tortues.txt", header = TRUE)
head(tortues)
```

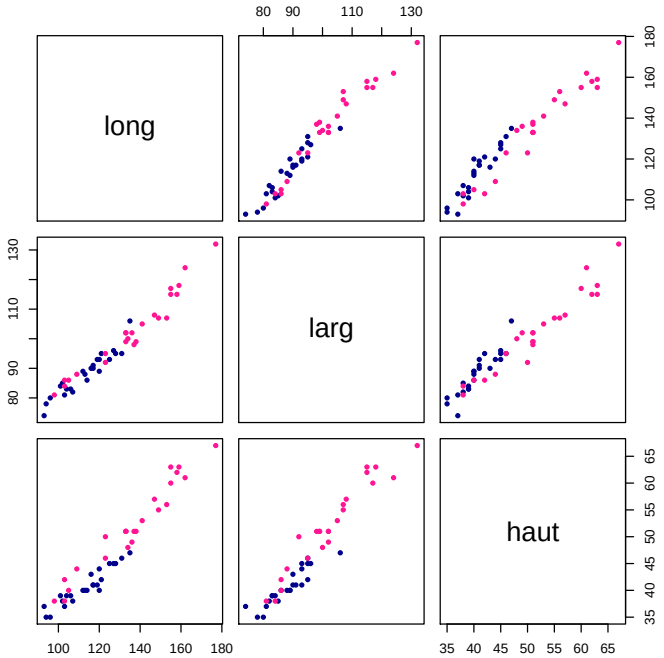
```
  long larg haut sexe
1   93   74   37    M
2   94   78   35    M
3   96   80   35    M
4  101   84   39    M
5  102   85   38    M
6  103   81   37    M
```

Trois variables sont mesurées : la hauteur, la longueur et la largeur de la carapace et sont toutes exprimées en millimètres. La quatrième concerne le sexe des tortues : mâle (`M`) ou femelle (`F`).
On peut représenter classiquement les tortues pour les trois variables prises deux à deux en y ajoutant l'information sexe (bleue pour les mâles et rose pour les femelles).

```
mesures <- tortues[,1:3]
dim(mesures)
```

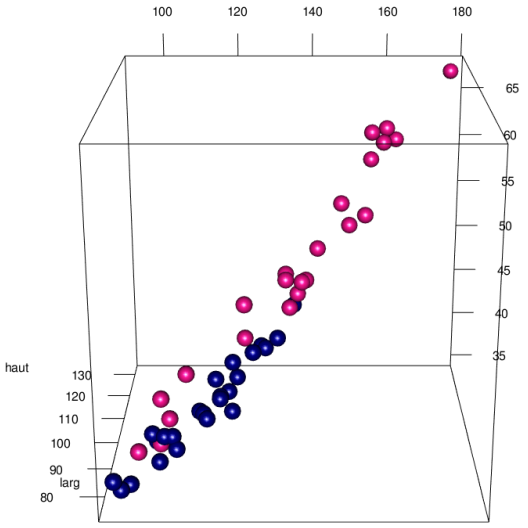
```
[1] 48  3
```

```
pairs(mesures, col=c("deeppink","blue4")[tortues$sexe], pch=20)
```



On peut également positionner les tortues dans l'espace des trois variables.

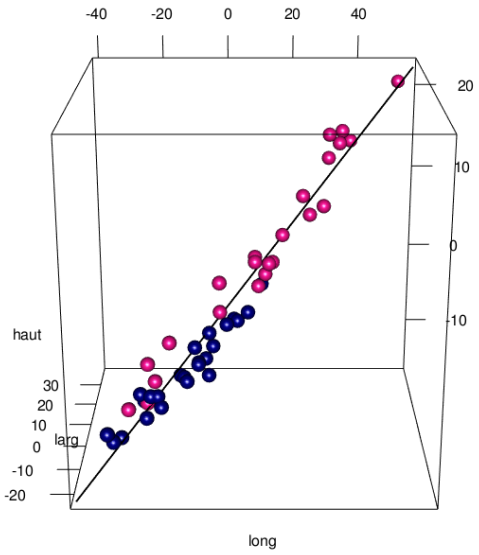
```
library(rgl)
plot3d(mesures, type = "s", col = c("deeppink","blue4")[tortues$sexe], size=1.5)
```



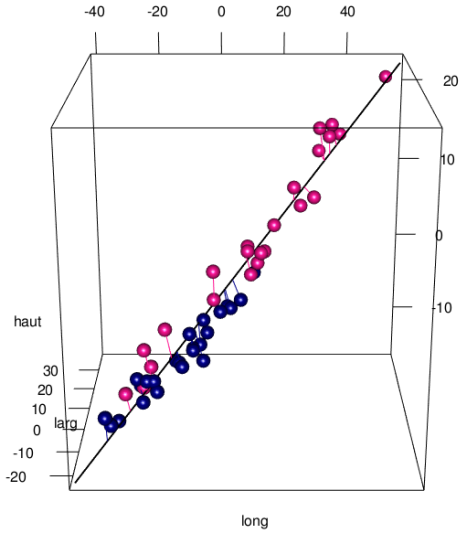
La question biologique est double.
• Peut-on faire une typologie des tortues ? comprendre celles qui se ressemblent, celles qui se différencient ? Quelles sont les variables responsables de cette typologie ?
• Peut-on faire une sélection des mesures effectuées sur les tortues ? Existe-t-il de la redondance entre les variables ? Certaines sont-elles plus judicieuses que d'autres pour décrire les tortues ?

La question mathématique est : comment faire pour répondre aux questions biologiques ? Quels outils mettre en place ? Concrètement, on veut représenter les lignes du tableau dans un espace de plus petite dimension en perdant le moins d'information possible. Peut-on, par exemple, visualiser les tortues dans un espace de dimension 2 tout en conservant un maximum de variabilité entre elles ? Peut-on visualiser les mesures effectuées sur ces tortues dans cet espace de dimension 2 et comprendre les liens qui les unissent ?

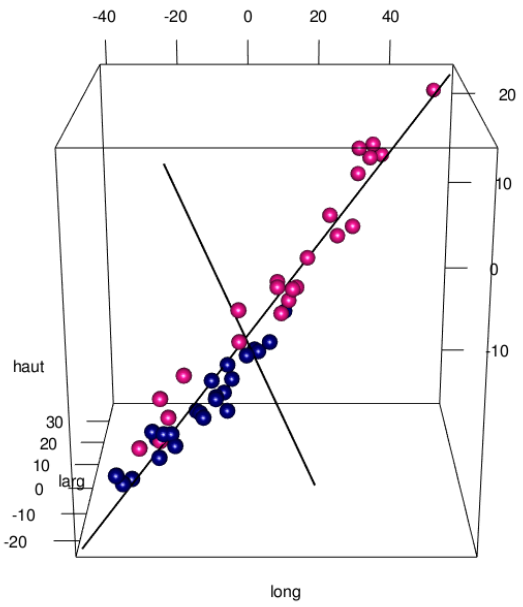
Le processus est le suivant. On recherche, sur les données centrées, une droite passant par l'origine :



telle que tous les points-tortues projetés orthogonalement sur cette droite soient de variance maximum.



Puis, on recherche le second axe, orthogonal au premier, tel que tous les points-tortues projetés orthogonalement sur ce dernier soient de seconde variance maximum.

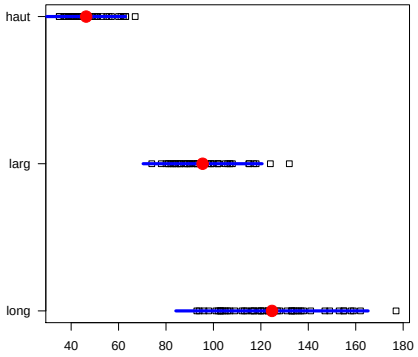


Dans ce type d'études, une question préalable se pose : vaut-il mieux travailler sur les données brutes, les données centrées ou les données centrées-réduites ? Pour ce faire, on construit une fonction sous R permettant de visualiser ces informations dont l'argument d'entrée est un data frame `df`.

```
graphdispersion <- function(df){
nbvar <- dim(df)[2]
moyennes <- sapply(df, mean)
variances <- sapply(df, var)*(dim(mesures)[1]-1)/dim(mesures)[1]
valsup <- max(moyennes+2*sqrt(variances))
valinf <- min(moyennes-2*sqrt(variances))
stripchart(df, las=1)
segments(moyennes-2*sqrt(variances),1:3,moyennes+2*sqrt(variances),1:3, lwd=4, col="blue")
points(moyennes,1:3,col="red",cex=3, pch=20)
}
```

On représente les trois mesures issues du tableau initial ce qu'on a appelé les données brutes.

```
graphdispersion(mesures)
```



On observe que les mesures, même si elles s'expriment dans les mêmes unités, sont en moyenne assez différentes. La variable "longueur" présente une grande dispersion. La variable "hauteur" présente une dispersion moindre. Deux tortues se distinguent par des valeurs nettement plus grandes que les autres.

Exercice.

1. On construit le tableau des données centrées :

```
mesures.c <- scalewt(mesures, center=TRUE, scale=FALSE)
dfmesures.c <- as.data.frame(mesures.c)
```

   Discuter des positions des moyennes et des dispersions des mesures.

2. Construire le tableau des données centrées réduites `dfmesures.cr` à l'aide de la fonction `scalewt`. Discuter des positions des moyennes et des dispersions des mesures.
3. Sur quelles données parait-il plus intéressant de travailler d'un point de vue biologique ? Justifier.

2 Des outils pour répondre à la question

Tout est affaire de statistique : moyennes, variances, covariances, corrélations mais aussi de relations dans des espaces : normes, distances, angles ... et donc de **produits scalaires**.

Point de vue 1.
On regarde les données par les **lignes** c'est-à-dire que chaque tortue est caractérisée par les trois variables : longueur, largeur et hauteur de la carapace. Les données d'une tortue sont stockées dans un vecteur de dimension 3.

Si on prend comme exemple la tortue 4, ces données correspondent à la quatrième ligne du tableau :

```
mesures[4,]
  long larg haut
4  101   84   39
```

On dit alors qu'elles sont stockées dans un vecteur $\begin{pmatrix} 101 \\ 84 \\ 39 \end{pmatrix}$.
Les $n = 48$ tortues sont regardées dans un espace de dimension 3 noté $\mathbb{R}^3$.

Point de vue 2.
On regarde les données par les **colonnes** c'est-à-dire que chaque variable est caractérisée par les 48 individus - tortues. Les données d'une variable sont stockées dans un vecteur de dimension 48 noté $\mathbb{R}^{48}$.

2.1 Définition du produit scalaire

Un vecteur de l'espace vectoriel $\mathbb{R}^n$ est un n-uplet de nombres réels :

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \quad \text{ou dans sa forme transposée } \mathbf{x}^\top = (x_1 \; x_2 \; \cdots \; x_n)$$

On définit $\omega_1, \omega_2, ..., \omega_n$, le poids de chacune des n lignes tel que $\sum_{i=1}^{n} \omega_i = 1$.

Soient deux vecteurs $\mathbf{x}$ et $\mathbf{y}$ de $\mathbb{R}^n$, le produit scalaire est défini par :

$$\langle \mathbf{x}|\mathbf{y} \rangle = \sum_{i=1}^{n} \omega_i x_i y_i.$$

Son écriture matricielle est donnée par $\mathbf{x}^\top \mathbf{D} \mathbf{y}$ où $\mathbf{D}$ est la matrice contenant les pondérations ω_i sur la première diagonale et des zéros partout ailleurs.

Définition
On dit que deux vecteurs sont orthogonaux si leur produit scalaire est nul.

Propriétés
- PS1 $\langle \mathbf{x}|\mathbf{y} \rangle = \langle \mathbf{y}|\mathbf{x} \rangle$ pour tout $\mathbf{x}$ et tout $\mathbf{y}$ de $\mathbb{R}^n$,
- PS2a $\langle \mathbf{x}|\mathbf{y} + \mathbf{z} \rangle = \langle \mathbf{x}|\mathbf{y} \rangle + \langle \mathbf{x}|\mathbf{z} \rangle$ pour tout $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$ de $\mathbb{R}^n$,
- PS2b $\langle \mathbf{x}|\alpha\mathbf{y} \rangle = \alpha \langle \mathbf{x}|\mathbf{y} \rangle$ pour tout $\mathbf{x}$ et $\mathbf{y}$ de $\mathbb{R}^n$ et pour tout α de $\mathbb{R}$,
- PS3 $\langle \mathbf{x}|\mathbf{x} \rangle \geq 0$ pour tout $\mathbf{x}$ de $\mathbb{R}^n$,
- PS4 $\langle \mathbf{x}|\mathbf{x} \rangle = 0 \Rightarrow \mathbf{x} = \mathbf{0}_n$ pour tout $\mathbf{x}$ de $\mathbb{R}^n$.

Le produit scalaire est une forme bilinéaire, symétrique, définie, positive.

Dans la suite du document, on pose $\forall i = 1, n \quad \omega_i = \frac{1}{n}$. Tous les individus ont le même poids et on parle de pondération uniforme. On fixe la valeur de n et on construit la fonction `prodscal` sous R :

```
n <- dim(mesures)[1]
prodscal <- function(x,y) (1/n)*sum(x*y)
```

Exercice.

1. Les données mesurées sont centrées (`mesures.c`).
   (a) Calculer le produit scalaire entre la longueur et la hauteur de la carapace.
   (b) Calculer la covariance entre la longueur et la hauteur de la carapace.
   (c) Comparer les deux résultats.
2. Les données mesurées sont centrées réduites (`mesures.cr`).
   (a) Calculer le produit scalaire entre la longueur et la hauteur de la carapace.
   (b) Calculer le coefficient de corrélation entre la longueur et la hauteur de la carapace.
   (c) Comparer les deux résultats.

On peut réaliser les calculs sur toutes les variables prises deux à deux et ceci de façon automatique grâce au calcul matriciel :

```
D <- diag(1/n, nrow=n , ncol=n)
(matps.c <- t(mesures.c) %*% D %*% mesures.c)
```

```
        long     larg      haut
long 410.7565 248.6992 162.37500
larg 248.6992 157.3294 100.06250
haut 162.3750 100.0625  68.97222
```

```
(matps.cr <- t(mesures.cr) %*% D %*% mesures.cr)
```

```
          long      larg      haut
long 1.0000000 0.9783116 0.9646946
larg 0.9783116 1.0000000 0.9605705
haut 0.9646946 0.9605705 1.0000000
```

Exercice.

1. Utiliser la fonction `cov` sur les données `mesures`. Correspond-elle à une des matrices précédentes ?
2. Utiliser la fonction `cor` sur les données `mesures`. Correspond-elle à une des matrices précédentes ?

2.2 Définition de la norme d'un vecteur

$$\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}|\mathbf{x}\rangle}.$$

Un vecteur normé $\mathbf{x}$ est défini par $\dfrac{\mathbf{x}}{\|\mathbf{x}\|}$.

On construit la fonction `norme` sous R :

```
norme <- function(x) sqrt(sum((1/n)*x*x))
```

Exercice
En utilisant la fonction `sapply`, calculer la norme de chacune des variables morphologiques centrées (`dfmesures.c`), centrées réduites (`dfmesures.cr`). Conclure.

2.3 Projection sur un vecteur

Si $\mathbf{x}$ et $\mathbf{u}$ sont deux vecteurs de $\mathbb{R}^n$ et si $\mathbf{u}$ est non nul, il existe un unique vecteur $\mathbf{z}$ de $\mathbb{R}^n$ proportionnel à $\mathbf{u}$ tel que $\mathbf{x} - \mathbf{z}$ soit orthogonal à $\mathbf{u}$.
On dit que $\mathbf{z}$ est le projeté orthogonal de $\mathbf{x}$ sur $\mathbf{u}$. Il vaut :

$$\mathbf{z} = \frac{\langle \mathbf{x}|\mathbf{u}\rangle}{\langle \mathbf{u}|\mathbf{u}\rangle}\ \mathbf{u}$$

Remarque.
Si $\mathbf{u}$ est normé, alors $\mathbf{z} = \langle \mathbf{x}|\mathbf{u}\rangle\ \ \mathbf{u}$.

Exercice.

1. Calculer la projection de la longueur de la carapace $\mathbf{x}$ sur le vecteur $\mathbf{1}_{48}$.
2. On note $\mathbf{z}$ le vecteur de cette projection. Ecrire le vecteur $\mathbf{x} - \mathbf{z}$. Calculer le carré de la norme. Que reconnait-on ?

2.4 Angle entre 2 vecteurs

Si $\mathbf{x}$ et $\mathbf{y}$ sont 2 vecteurs-points de $\mathbb{R}^n$, la mesure de l'angle de $\mathbf{x}$ et $\mathbf{y}$ est notée

$$A\left(\mathbf{x}, \mathbf{y}\right) = a$$

Elle est définie par :

$$0 \leq a \leq \pi \text{ et } \cos a = \frac{\langle \mathbf{x}|\mathbf{y}\rangle}{\|\mathbf{x}\|\|\mathbf{y}\|}.$$

Exercice.

1. Calculer le cosinus de l'angle entre la longueur et les hauteur et largeur de la carapace.
2. Calculer le cosinus de l'angle entre la longueur et les hauteur et largeur de la carapace, variables centrées.
3. Calculer le cosinus de l'angle entre la longueur et les hauteur et largeur de la carapace, variables centrées réduites.
4. Comparer ces résultats aux coefficients de corrélation. Que peut-on en déduire ?

2.5 Distance entre 2 vecteurs

On lit le tableau par les lignes et on calcule des normes et des distances sur les tortues mais la pondération n'est plus la même. On associe 1 à chacune des trois variables. On parle alors de pondération canonique.

Les définitions du produit scalaire et de la norme deviennent :

$$\langle \mathbf{x}|\mathbf{y}\rangle = \sum_{i=1}^{p} x_i y_i$$

$$\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}|\mathbf{x}\rangle}$$

et la distance entre deux lignes-tortues :

$$d(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\| = \|\mathbf{y}-\mathbf{x}\|.$$

La distance ainsi calculée est appelée **distance euclidienne**. On construit sa fonction associée `distance` sous R :

```
neonorme <- function(x) sqrt(sum(x*x))
distance <- function(x,y) neonorme(x-y)
```

Exercice.

1. Calculer les distances entre les tortues 4, 5 et 6. Décrire les relations de proximité entre les trois tortues.
2. La fonction `dist` de R permet de réaliser ces calculs automatiquement. On construit la matrice des distances entre les 48 tortues.

   ```
   resdist <- dist(mesures,diag=TRUE,upper=TRUE)
   matdist <- as.matrix(resdist)
   ```

   Afficher les lignes 4, 5, 6 et les colonnes 4, 5, 6 de la matrice de distances. Vérifier que les résultats trouvés sont les mêmes qu'à la question 1.

3 Résolution du problème

L'objectif est de rechercher des combinaisons linéaires des variables appelées axes principaux qui maximisent la variance de la projection des points sur ces axes.

La clef du problème est la diagonalisation c'est-à-dire la recherche de vecteurs propres. On dit que $\mathbf{v}$ est vecteur propre si et seulement si :

1. $\mathbf{v} \neq 0$
2. $\exists \lambda \in \mathbb{R}, \qquad \mathbf{X}\mathbf{v} = \lambda \mathbf{v}$

Le scalaire λ est dit **valeur propre** associée à $\mathbf{v}$. Sous R, il existe une procédure automatique de recherche des valeurs et vecteurs propres : la fonction `eigen`. On note $\mathbf{X}$ la matrice associée aux trois mesures réalisées sur les carapaces des tortues. On note $\mathbf{X}_0$ cette matrice quand les données sont centrées. Elles sont toutes les deux de dimension 48×3. On définit :

```
matX0 <- mesures.c
```

3.1 48 points dans un espace de dimension 3

On regarde les données par les lignes.

1. En utilisant la commande `%*%` qui permet de calculer le produit de deux matrices, construire la matrice `matC` : $\mathbf{C} = \frac{1}{n}\,\mathbf{X}_0{}^\top \mathbf{X}_0$. Donner ces dimensions.
2. Comparer le résultat à `cov(mesures)`. Conclure.
3. Rechercher les valeurs propres et les vecteurs propres de $\mathbf{C}$. Les vecteurs propres seront stockés dans une matrice $\mathbf{U}$ qu'on écrira `matU` sous R.

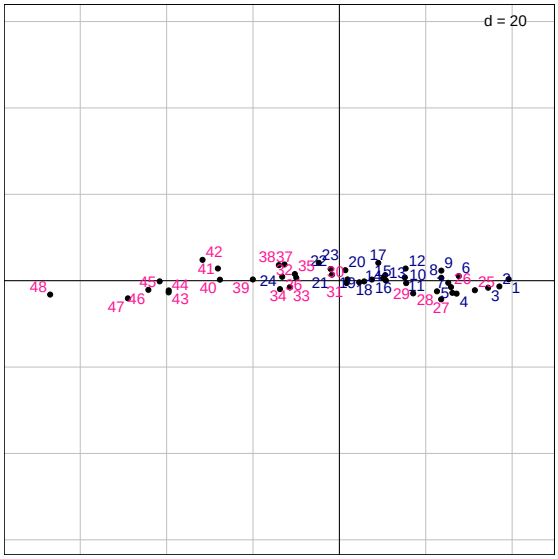
4. Calculer les coordonnées des lignes-tortues dans l'espace des vecteurs propres : $\mathbf{L} = \mathbf{X}_0\mathbf{U}$ et les stocker dans `matL`.
5. On calcule les variances descriptives des coordonnées des tortues dans l'espace des vecteurs propres.

```
sapply(as.data.frame(matL), var)*(n-1)/n
```

```
        V1         V2         V3
628.211335   5.095961   3.750864
```

   Comparer les résultats aux valeurs propres. Conclure.
6. On représente les tortues dans le plan engendré par les deux premiers vecteurs propres. Commenter.

```
library(adegraphics)
dfL <- as.data.frame(matL)
s.label(dfL, plabel.col=couleur, plabel.boxes.draw=FALSE, plabel.optim=TRUE)
```



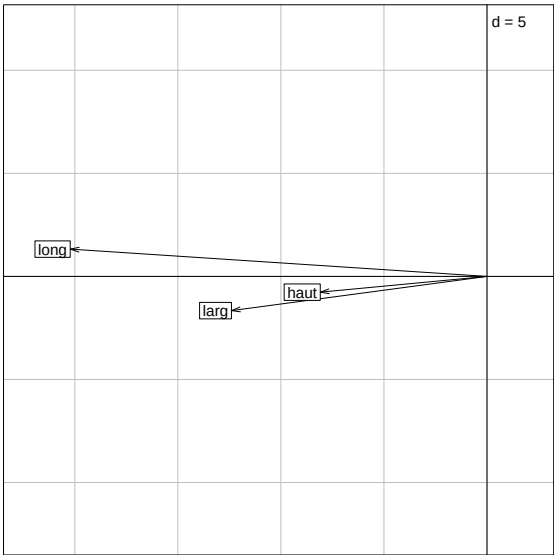
3.2 3 points dans un espace de dimension 48

On regarde les données par les colonnes.

1. En utilisant la commande `%*%`, construire la matrice `matS` : $\mathbf{S} = \frac{1}{n}\,\mathbf{X}_0\,\mathbf{X}_0^\top$. Donner ces dimensions.
2. Rechercher les valeurs propres et les vecteurs propres de $\mathbf{S}$. Les vecteurs propres seront stockés dans une matrice $\mathbf{V}$ qu'on écrira `matV` sous R. Comparer les valeurs propres non nulles de $\mathbf{C}$ et de $\mathbf{S}$.
3. Représenter les coordonnées des individus sur l'axe 1 (`matL[,1]`) et le premier vecteur propre $\mathbf{V}$ issu de la diagonalisation de $\mathbf{S}$ (`matV[,1]`). Qu'observe-t-on ? Qu'est-ce que cela signifie ?
4. Calculer la norme de `matL[,1]` à l'aide de la fonction `norme` ; la norme de `matV[,1]` à l'aide de la fonction `neonorme`. Comparer les résultats obtenus à la racine carrée de la première valeur propre `sqrt(resS$values[1])`. Ecrire la relation liant les deux vecteurs.

5. Compte-tenu de ce qui précède, on peut représenter les trois variables dans le plan des vecteurs propres $\mathbf{U}$: $\mathbf{K} = \mathbf{U}\mathbf{\Lambda}^{1/2}$ où $\mathbf{\Lambda}$ est la matrice contenant les valeurs propres non nulles sur sa diagonale et des 0 partout ailleurs.

```
matK <- matU%*%diag(sqrt(resC$values))
s.arrow(matK, labels=colnames(mesures))
```

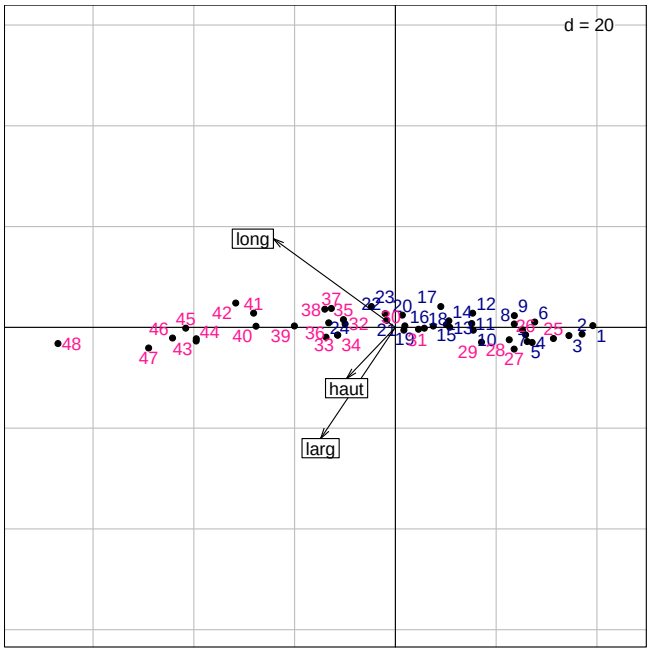


La projection orthogonale de la longueur, de la largeur et de la hauteur de la carapace sur l'axe horizontal (puis sur l'axe vertical) définit le sens biologique de l'axe.

3.3 Représentation simultanée

L'analyse de $n = 48$ points dans l'espace de dimension $p = 3$ et l'analyse de $p = 3$ points dans l'espace de dimension $n = 48$ ont montré que toutes deux étaient liées. Par conséquent, on a, à la fois, la compréhension des axes principaux, combinaisons linéaires des variables de départ, et la compréhension de la typologie (ressemblances et différences) des individus.
On peut donc représenter sur un même graphique les deux informations. On parle de **biplot**. L'interprétation qui s'en suit est la même que dans les paragraphes précédents.

```
s.label(dfL, plabel.col=couleur, plabel.boxes.draw=FALSE, plabel.optim=TRUE)
s.arrow(30*matU, labels=colnames(mesures), col="red", add=TRUE)
```



4 Conclusion

L'étude que l'on vient de réaliser s'appelle une **analyse en composantes principales** (ACP). Ses objectifs sont
(1) la typologie des individus, point de vue que l'on doit à K. Pearson (1901) [3]
(2) la sélection de variables, point de vue que l'on doit à Hotelling (1933) [1].
C'est la plus ancienne analyse multivariée connue. Elle est très utilisée en écologie. L'ACP permet d'étudier l'abondance d'espèces sur des sites échantillonnés afin de comprendre leur organisation. L'ACP permet de caractériser des sites en fonction de catactéristiques physico-chimiques afin de repérer par exemple des zones de pollution.
Enfin, cette méthode est la base de nombreux développements méthodologiques : données fonctionnelles, couplage avec des données spatialisées, etc.

Références

[1] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24 :498–520, 1933.

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

[3] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Transactions of the Royal Society London Series A - Physics*, 6(2) :559–572, 1901.