Structures morphologiques sur les Crabs

Yéro Diamanka

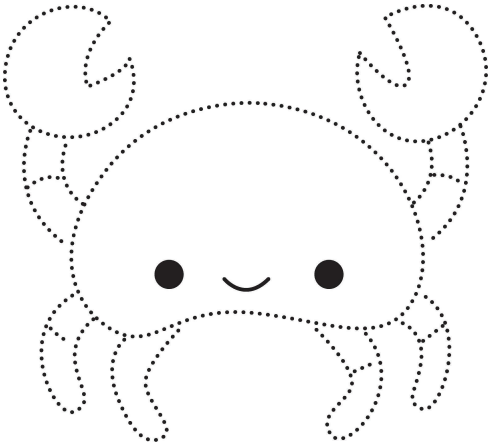
Rapport de Résulats

Table des matières

1 Introduction

Les crabes de l'espèce *Leptograpsus variegatus*, collectés à Fremantle en Australie-Occidentale, se déclinent en deux formes de couleur (bleue et orange) et deux sexes (mâle et femelle), formant quatre groupes distincts de 50 individus chacun. Cette espèce, caractéristique des côtes australiennes, présente une grande variabilité morphologique dont l'étude permet de mieux comprendre les mécanismes de différenciation entre formes et entre sexes.

Cinq mesures morphologiques ont été relevées sur 200 individus : la taille du lobe frontal (FL), la largeur postérieure (RW), la longueur (CL) et la largeur (CW) de la carapace, ainsi que la profondeur du corps (BD). Ces variables, toutes exprimées en millimètres, présentent de fortes corrélations entre elles, ce qui motive le recours à une méthode de réduction de dimension.



L'objectif de ce travail est de réaliser une **Analyse en Composantes Principales(ACP)** des données `crabs` du package MASS en R afin d'identifier les structures morphologiques sous-jacentes, de synthétiser l'information portée par ces cinq variables et d'explorer les différences morphologiques entre espèces et sexes.

2 Visualisation des données

On peut installer la bibliothèque MASS en R et charger les données "crabs".

```
# Charger la bibliothéque
library(MASS)
```

Nous allons charger les données et afficher les 10 premières lignes. Pour afficher les toutes les données, vous pouvez faire `crabs` au lieu de `head(crabs, 10)`.

```
# Charger les données crabs
data(crabs)

# Afficher les premières lignes
head(crabs, 10)
```

```
   sp sex index   FL   RW   CL   CW   BD
1   B   M     1  8.1  6.7 16.1 19.0  7.0
2   B   M     2  8.8  7.7 18.1 20.8  7.4
3   B   M     3  9.2  7.8 19.0 22.4  7.7
4   B   M     4  9.6  7.9 20.1 23.1  8.2
5   B   M     5  9.8  8.0 20.3 23.0  8.2
6   B   M     6 10.8  9.0 23.0 26.5  9.8
7   B   M     7 11.1  9.9 23.8 27.1  9.8
8   B   M     8 11.6  9.1 24.5 28.4 10.4
9   B   M     9 11.8  9.6 24.2 27.8  9.7
10  B   M    10 11.8 10.5 25.2 29.3 10.3
```

```
summary(crabs)
```

```
 sp      sex         index            FL              RW              CL
 B:100   F:100   Min.   : 1.0   Min.   : 7.20   Min.   : 6.50   Min.   :14.70
 O:100   M:100   1st Qu.:13.0   1st Qu.:12.90   1st Qu.:11.00   1st Qu.:27.27
                 Median :25.5   Median :15.55   Median :12.80   Median :32.10
                 Mean   :25.5   Mean   :15.58   Mean   :12.74   Mean   :32.11
                 3rd Qu.:38.0   3rd Qu.:18.05   3rd Qu.:14.30   3rd Qu.:37.23
                 Max.   :50.0   Max.   :23.10   Max.   :20.20   Max.   :47.60
       CW              BD
 Min.   :17.10   Min.   : 6.10
 1st Qu.:31.50   1st Qu.:11.40
 Median :36.80   Median :13.90
 Mean   :36.41   Mean   :14.03
 3rd Qu.:42.00   3rd Qu.:16.60
 Max.   :54.60   Max.   :21.60
```

```
# Afficher le nombre d'observations
nrow(crabs)
```

```
[1] 200
```

Le jeu de données crabs comporte 200 lignes et 8 colonnes, décrivant 5 mesures morphologiques sur 50 crabes de chacune des deux formes de couleur et des deux sexes, de l'espèce Leptograpsus variegatus collectés à Fremantle, en Australie-Occidentale.

Format

Une trame de données avec 200 observations sur les 6 variables suivantes.

`classe` : Type de crabes : le premier caractère représente l'espèce « B » ou « O » pour bleu ou orange, le second représente le sexe « M » ou « F » pour mâle ou femelle.

–> `FL` : Taille du lobe frontal (mm).

–> `RW` : Largeur arrière (mm).

–> `CL` : Longueur de la carapace (mm).

–> `CW` : largeur de la carapace (mm).

–> `BD` : Profondeur du corps (mm).

3 Préliminaires (statistiques et corrélations)

```
library(FactoMineR)
```

```
Warning: package 'FactoMineR' was built under R version 4.5.3
```

```
# Extraire les variables quantitatives
crabs_qt <- crabs[,4:8]
```

```
Moyenne des variables : colMeans(crabs_qt)
```

```
     FL      RW      CL      CW      BD
15.5830 12.7385 32.1055 36.4145 14.0305
```

```
Matrice de corrélation : cor(crabs_qt)
```

```
          FL        RW        CL        CW        BD
FL 1.0000000 0.9069876 0.9788418 0.9649558 0.9876272
RW 0.9069876 1.0000000 0.8927430 0.9004021 0.8892054
CL 0.9788418 0.8927430 1.0000000 0.9950225 0.9832038
CW 0.9649558 0.9004021 0.9950225 1.0000000 0.9678117
BD 0.9876272 0.8892054 0.9832038 0.9678117 1.0000000
```

La matrice de corrélation révèle des corrélations très fortes (proches de 1) entre toutes les variables morphologiques, en particulier entre `CL` et `CW` (longueur et largeur de la carapace), et entre `FL` et `CL`. La variable `RW` (largeur de la région postérieure) est un peu moins corrélée aux autres, mais reste très liée à l'ensemble. Ces corrélations élevées suggèrent l'existence d'un fort **effet taille** commun à toutes les variables, que l'ACP permettra de mettre en évidence.

Rappelons que dans le cours, la matrice de corrélation s'écrit $R = DSD = Z^\top PZ$, où Z est le tableau centré-réduit et $P = \frac{1}{n} I$ la matrice des poids des individus.

4 ACP des données

Bien que les variables soient toutes exprimées en millimètres, nous réalisons une **ACP normée** (`scale.unit = TRUE`). En effet, même si les unités sont homogènes, les variables présentent des variances différentes : les variables les plus dispersées (comme `CL` et `CW`) risqueraient de dominer l'analyse si on ne réduisait pas. L'ACP normée revient à diagonaliser la matrice de corrélation R, ce qui donne le même poids à chaque variable.

```
# Réaliser l'ACP
res_pca <- PCA(crabs[, 4:8], scale.unit = TRUE, ncp = 5, graph = FALSE)
```

4.1 Commentaires sur les valeurs propres et les inerties

Les valeurs propres et les inerties permettent de déterminer le nombre d'axes principaux à retenir et de mesurer la dispersion totale du nuage de points. Les valeurs propres indiquent la direction de la plus grande variance dans les données, tandis que les inerties mesurent la quantité de variance dans ces directions.

```
# Afficher les valeurs propres et inertie
print(res_pca$eig)
```

```
       eigenvalue percentage of variance cumulative percentage of variance
comp 1 4.788834784             95.77669569                           95.77670
comp 2 0.151685207              3.03370413                           98.81040
comp 3 0.046632974              0.93265948                           99.74306
comp 4 0.011135357              0.22270714                           99.96577
comp 5 0.001711678              0.03423355                          100.00000
```

En théorie, l'inertie totale du nuage des individus vaut $I_g = \text{Trace}(MS)$. Pour l'ACP normée ($M = D^2$), on a $I_g = \text{Trace}(R) = p = 5$.

La somme des valeurs propres est donc égale à 5, et chaque valeur propre représente la part d'inertie captée par l'axe correspondant :

–> Les deux premiers axes principaux expliquent 98.81% de la variance totale (PC1: 95.77%, cumulée: 95.77%).

–> Le premier axe (PC1) capture 95.77.96% de l'inertie, ce qui est très élevé.

–> Les deux premiers axes suffisent à expliquer la majorité de la variabilité des données.

–> Les 3ème, 4ème et 5ième axes contribuent peu (< 1,2% cumulé jusqu'aux 5 premiers axes représentent 100%).

On distigue en effet de plusieurs de **critères de sélection des axes**, parmi ces critères, on a :

–> *Le critère du coude* : le graphique des valeurs propres montre un coude très marqué après le 1er axe.

–> *Le critère de Kaiser* : pour l'ACP normée, on retient les axes dont la valeur propre est supérieure à 1. Seul le 1er axe satisfait ce critère.

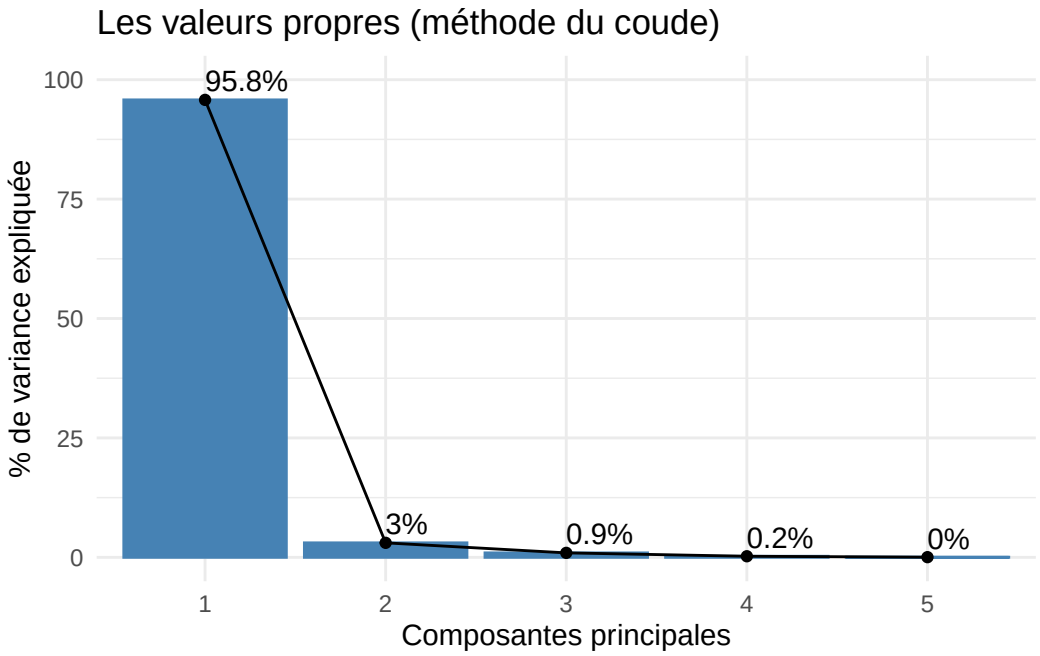
On se limitera donc à l'interprétation des **deux premiers axes**, qui suffisent à décrire la structure essentielle des données.

```
library(factoextra)
```

```
Warning: package 'factoextra' was built under R version 4.5.3
```

```
Warning: package 'ggplot2' was built under R version 4.5.3
```

```
# Les valeurs propres (méthode du coude)
fviz_eig(res_pca, addlabels = TRUE, ylim = c(0, 100)) +
  labs(title = "Les valeurs propres (méthode du coude)",
       x = "Composantes principales", y = "% de variance expliquée") +
  theme_minimal()
```



4.2 Commentaires sur l'ACP des variables

–> Résultats sur l'**acp** des variables : `res_pca$var`

```
$coord
       Dim.1       Dim.2       Dim.3        Dim.4       Dim.5
FL 0.9892256 -0.05358348 -0.114617683 -0.0735422825  0.003992090
RW 0.9367791  0.34979304  0.002586857  0.0088327137 -0.002251388
CL 0.9917363 -0.10447013  0.066874707  0.0001524437 -0.032753833
CW 0.9871883 -0.07033629  0.140920239 -0.0094114731  0.023769564
BD 0.9872340 -0.10294487 -0.095699159  0.0745672444  0.007270927

$cor
       Dim.1       Dim.2       Dim.3        Dim.4       Dim.5
FL 0.9892256 -0.05358348 -0.114617683 -0.0735422825  0.003992090
RW 0.9367791  0.34979304  0.002586857  0.0088327137 -0.002251388
CL 0.9917363 -0.10447013  0.066874707  0.0001524437 -0.032753833
CW 0.9871883 -0.07033629  0.140920239 -0.0094114731  0.023769564
BD 0.9872340 -0.10294487 -0.095699159  0.0745672444  0.007270927

$cos2
       Dim.1       Dim.2        Dim.3        Dim.4        Dim.5
FL 0.9785672 0.002871189 1.313721e-02 5.408467e-03 1.593678e-05
RW 0.8775551 0.122355169 6.691831e-06 7.801683e-05 5.068747e-06
CL 0.9835409 0.010914009 4.472226e-03 2.323908e-08 1.072814e-03
CW 0.9745407 0.004947193 1.985851e-02 8.857583e-05 5.649922e-04
BD 0.9746309 0.010597646 9.158329e-03 5.560274e-03 5.286638e-05

$contrib
      Dim.1     Dim.2     Dim.3        Dim.4      Dim.5
FL 20.43435  1.892860 28.171510 4.857022e+01  0.9310620
```

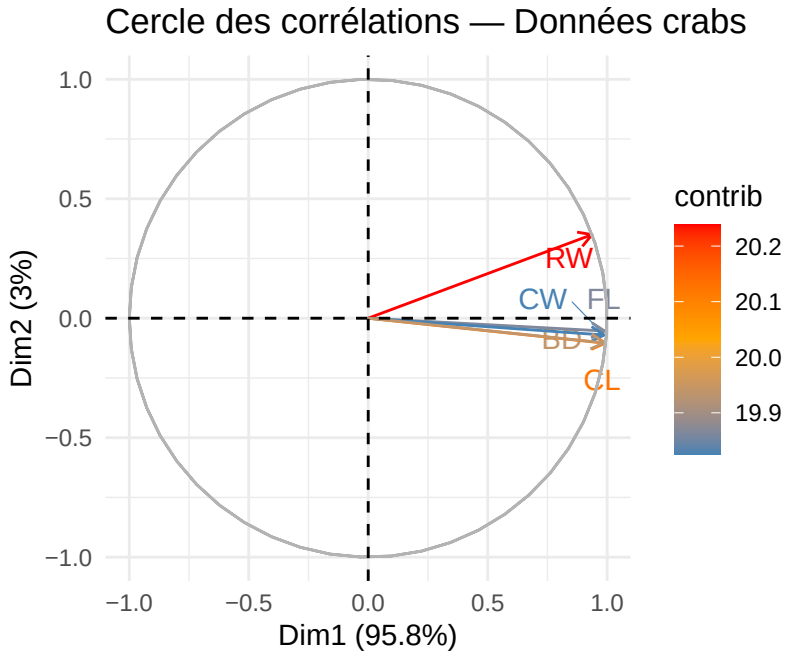
```
RW 18.32502 80.663877  0.014350 7.006226e-01  0.2961274
CL 20.53821  7.195170  9.590266 2.086963e-04 62.6761450
CW 20.35027  3.261487 42.584703 7.954467e-01 33.0080946
BD 20.35215  6.986605 19.639170 4.993350e+01  3.0885710
```

–> **Axe 1 :** Toutes les variables (`FL`, `RW`, `CL`, `CW`, `BD`) sont très bien représentées sur le 1er axe (cos^2 proches de 1) et ont des coordonnées positives élevées. Elles contribuent toutes fortement à sa construction. Cet axe est un **axe de taille** : les individus à droite sont les plus grands (toutes mesures morphologiques grandes), ceux à gauche sont les plus petits.

–> **Axe 2 :** La variable `RW` (largeur postérieure) se distingue des autres : elle est mieux représentée sur le 2ème axe et lui est positivement corrélée, tandis que les variables `CL`, `CW`, `FL` et `BD` sont négativement corrélées avec cet axe. Cet axe oppose donc `RW` aux autres variables et constitue un **axe de forme** ou de **proportion corporelle**.

Les autres axes ne sont pas interprétables (faible inertie, mauvaise représentation des variables).

```r
# Cercle des corrélations - plan (PC1, PC2)
fviz_pca_var(res_pca,
             col.var = "contrib",
             gradient.cols = c("steelblue", "orange", "red"),
             repel = TRUE) +
  labs(title = "Cercle des corrélations - Données crabs") +
  theme_minimal()
```



4.3 L'ACP des individus

–> Résultats des sur l'**acp** des individus : `head(round(res_pca$ind$coord, 3), 10)`

```
    Dim.1  Dim.2  Dim.3  Dim.4  Dim.5
1  -4.928 -0.268 -0.122  0.039  0.069
```

```
2  -4.386 -0.094 -0.039 -0.005 -0.003
3  -4.129 -0.169  0.034 -0.038  0.038
4  -3.884 -0.246  0.015 -0.019  0.001
5  -3.834 -0.224 -0.015 -0.055 -0.025
6  -2.953 -0.220  0.038  0.070  0.019
7  -2.678  0.039  0.082  0.033 -0.037
8  -2.548 -0.363  0.063  0.016  0.042
9  -2.585 -0.117  0.062 -0.146 -0.010
10 -2.206  0.079  0.157 -0.009  0.000
```

–> Sur le **1er axe**, les individus se distribuent selon leur taille globale, sans distinction immédiate d'espèce ou de sexe. Les individus aux coordonnées positives élevées correspondent aux crabes les plus grands (grandes valeurs de toutes les variables morphologiques).

–> Sur le **2ème axe**, une structure plus fine apparaît, qui pourrait correspondre à des différences de morphologie entre les espèces ou les sexes. Les individus avec des coordonnées positives sur cet axe ont une largeur postérieure (`RW`) relativement grande par rapport à leur taille, tandis que ceux avec des coordonnées négatives ont des carapaces relativement plus longues et larges.

Centres des classes : `res_pca$call$centre`

```
[1] 15.5830 12.7385 32.1055 36.4145 14.0305
```

Le centre de gravité g (vecteur des moyennes) correspond à l'origine du repère factoriel, conformément aux propriétés de l'ACP : les composantes principales sont des variables centrées.

5 Produire les graphiques

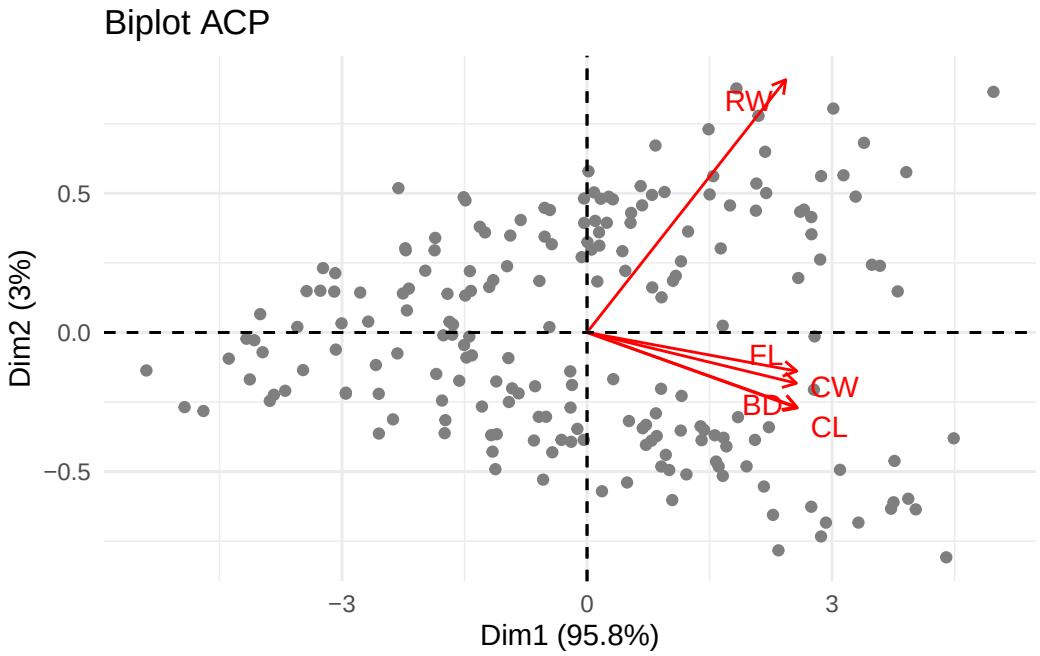
5.1 Nuage des individus et cercle des corrélations

```
res_pca <- PCA(crabs_qt, scale.unit = TRUE, ncp = 5, graph = TRUE)
```

5.2 Le biplot

Le biplot superpose le nuage des individus et le cercle des corrélations dans le même repère factoriel.

```
fviz_pca_biplot(res_pca,
                label = "var",
                col.ind = "grey50",
                col.var = "red",
                repel = TRUE) +
  labs(title = "Biplot ACP") +
  theme_minimal()
```



–> Les individus situés à droite ont de grandes valeurs pour toutes les variables morphologiques (grands crabes).

–> La variable `RW` pointe légèrement vers le haut, tandis que les autres variables pointent vers la droite : les individus situés en haut ont un rapport `RW`/taille globale plus élevé.

6 Analyse des résultats dans les plans factoriels

Analyse dans le 1er plan factoriel

–> Toutes les variables sont bien représentées dans ce plan factoriel (toutes sont proches du bord du cercle des corrélations).

–> Les variables `CL`, `CW`, `FL` et `BD` sont très fortement corrélées entre elles et toutes positivement associées à l'axe 1.

–> La variable `RW` est légèrement décalée des autres, avec une composante positive sur l'axe 2.

–> On retrouve ici la structure de la matrice de corrélation : corrélations très élevées pour la plupart des paires, légèrement plus faibles pour `RW`.

Ce plan révèle que les individus se différencient surtout par leur **taille globale** (axe 1), et plus subtilement par leurs **proportions morphologiques** (axe 2), notamment le rapport entre la largeur postérieure et la taille de la carapace.

7 Variables et individus supplémentaires

7.1 Variables supplémentaires qualitatives : `sp` et `sex`

La variable d'espèce (`sp`) et la variable de sexe (`sex`) peuvent être introduites comme variables **supplémentaires qualitatives** pour interpréter les axes.

```
# Création d'une variable combinée espèce × sexe
crabs$groupe <- as.factor(paste(crabs$sp, crabs$sex, sep = "_"))
res_pca_indsup <-PCA(crabs[, 4:9], scale.unit= TRUE,ncp= 4,
                ind.sup=c(10, 55, 120,175),
                quali.sup= 6,graph=TRUE)
```

Résultats de la variable qualitative supplémentaire : `res_pca_indsup$quali.sup`

```
$coord
         Dim.1      Dim.2      Dim.3       Dim.4
B_F -1.1385534  0.2983918  0.1492784  0.02757172
B_M -0.3299691 -0.2888072  0.2252253 -0.03092967
O_F  1.1043997  0.3776703 -0.1782043 -0.03125326
O_M  0.3641228 -0.3872549 -0.1962995  0.03461121

$cos2
        Dim.1      Dim.2      Dim.3        Dim.4
B_F 0.9204063 0.06321885 0.01582220 0.0005397598
B_M 0.4462317 0.34184546 0.20789704 0.0039207063
O_F 0.8742769 0.10224029 0.02276318 0.0007001434
O_M 0.4111242 0.46501948 0.11948558 0.0037145893

$v.test
        Dim.1     Dim.2     Dim.3     Dim.4
B_F -4.196000  6.125788  5.551561  2.091358
B_M -1.216061 -5.929023  8.375974 -2.346064
O_F  4.070131  7.753324 -6.627295 -2.370608
O_M  1.341930 -7.950089 -7.300240  2.625314

$dist
      B_F       B_M       O_F       O_M
1.1867619 0.4939615 1.1811411 0.5678862

$eta2
          Dim.1     Dim.2     Dim.3      Dim.4
groupe 0.144046 0.7538328 0.7622742 0.08611481
```

–> Le **1er axe** ne discrimine pas bien les groupes (espèce × sexe) : les quatre groupes (B_F, B_M, O_F, O_M) ont des coordonnées proches sur cet axe, ce qui confirme que l'axe 1 capte principalement l'**effet taille** et non les différences entre groupes.

–> Le **2ème axe** sépare davantage les groupes. Il tend à discriminer les espèces bleue (B) et orange (O), suggérant des différences de **forme** (notamment dans les proportions de `RW`) entre les deux espèces.

Les valeurs-test permettent d'identifier les positions significativement différentes de la moyenne générale ($|\text{valeur-test}| > 2$).

Distances des groupes au centre de gravité :

```
# Moyennes par groupe
means <- aggregate(crabs[, 4:8], by = list(crabs$groupe), FUN = mean)
print(means)
```

```
  Group.1     FL     RW     CL     CW     BD
1     B_F 13.270 12.138 28.102 32.624 11.816
2     B_M 14.842 11.718 32.014 36.810 13.350
3     O_F 17.594 14.836 34.618 39.036 15.632
4     O_M 16.626 12.262 33.688 37.188 15.324
```

Centre de gravité général : `colMeans(crabs[, 4:8]`

```
     FL      RW      CL      CW      BD
15.5830 12.7385 32.1055 36.4145 14.0305
```

```r
# Distances euclidiennes au centre
dist_centre <- apply(means[, -1], 1, function(x) sqrt(sum((x - centre)^2)))
names(dist_centre) <- means[, 1]
```

Distances euclidiennes au centre de gravité : `dist_centre`

```
     B_F      B_M      O_F      O_M
6.403943 1.489421 4.918673 2.467925
```

7.2 Individus supplémentaires

Il est possible d'introduire certains individus comme **supplémentaires** pour vérifier leur positionnement dans le plan factoriel sans qu'ils participent à la construction des axes.

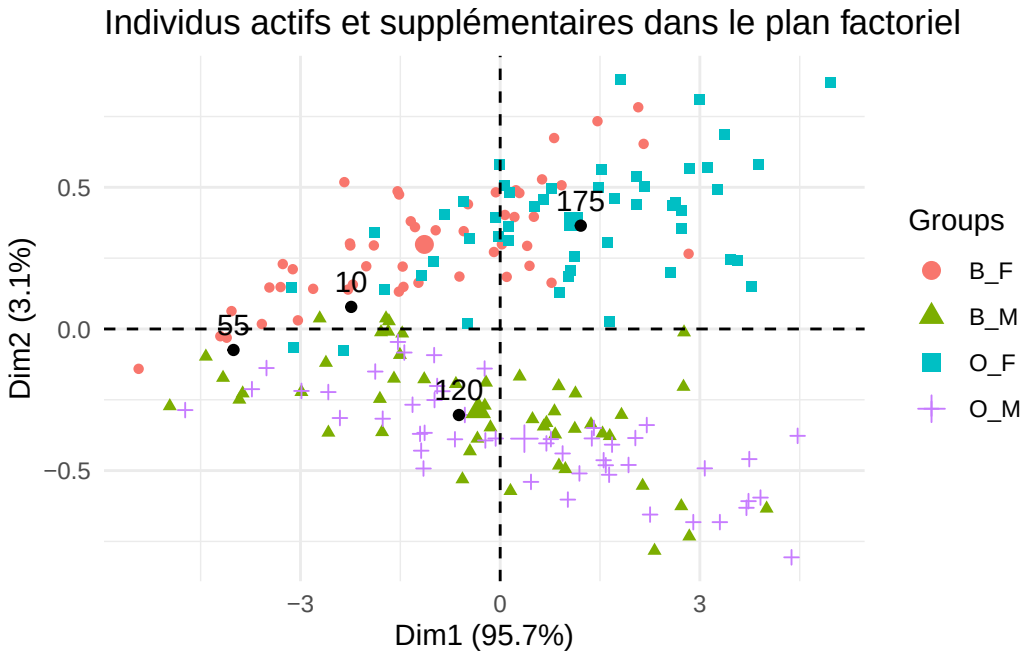
```r
# ACP avec individus supplémentaires (un représentant de chaque groupe)
res_pca_indsup <- PCA(crabs[, 4:9], scale.unit = TRUE, ncp = 4,
ind.sup = c(10, 55, 120, 175),
quali.sup = 6, graph = FALSE)
```

Coordonnées des individus supplémentaires : `round(res_pca_indsup$ind.sup$coord, 3)`

```
     Dim.1  Dim.2  Dim.3  Dim.4
10  -2.238  0.078  0.157 -0.010
55  -4.006 -0.074  0.003 -0.065
120 -0.619 -0.304 -0.273  0.082
175  1.211  0.365 -0.207 -0.103
```

Visualisation avec individus supplémentaires

```r
fviz_pca_ind(res_pca_indsup,
             habillage = crabs[-c(10, 55, 120, 175), "groupe"],
             addEllipses = FALSE,
             label = "ind.sup",
             col.ind.sup = "black",
             pointsize = 1.5) +
  labs(title = "Individus actifs et supplémentaires dans le plan factoriel") +
  theme_minimal()
```



–> Les individus supplémentaires (indices 10, 55, 120, 175, représentant les quatre groupes espèce × sexe) se positionnent de manière cohérente dans le plan factoriel, à des emplacements proches des autres individus de leur groupe, ce qui valide la robustesse de l'analyse.

Conclusion :

–> Cette analyse en composantes principales des données *crabs* a permis de mettre en évidence la structure morphologique de 200 crabes de l'espèce *Leptograpsus variegatus*. Les résultats montrent que deux axes principaux suffisent à expliquer 98,81 % de la variance totale, ce qui témoigne d'une structure très forte dans les données.

–> Le premier axe, qui capture à lui seul 95,77 % de l'inertie, représente un **effet taille** global : toutes les variables morphologiques (FL, RW, CL, CW, BD) y contribuent positivement et de manière quasi équivalente. Le second axe, bien que marginal en termes d'inertie (3,04 %), révèle un **effet de forme** en opposant la largeur postérieure (RW) aux autres mesures de la carapace, ce qui traduit des différences de proportions corporelles entre individus.

–> L'introduction des variables qualitatives supplémentaires (espèce et sexe) a confirmé que l'axe 1 ne discrimine pas les groupes, tandis que l'axe 2 tend à séparer les espèces bleue et orange, suggérant des différences morphologiques subtiles entre elles. Les individus supplémentaires se sont positionnés de manière cohérente avec leur groupe d'appartenance, validant ainsi la robustesse de l'analyse.

–> En définitive, l'ACP normée s'est révélée particulièrement adaptée à ce jeu de données, permettant de résumer efficacement une information multidimensionnelle en un plan factoriel lisible et interprétable.