Parcial 3

Puntos 3 y 4

Cargamos los datos. Las variables independientes son las 8 primeras columnas.

```
library(MASS)
load("abulon.RData")
head(abulon) # El dataset es la variable abulon.
```

```
##    Sexo Longitud Diametro Altura Peso_total Peso_sin_concha Peso_visceras
## 3     1    0.530    0.420  0.135     0.6770          0.2565        0.1415
## 4     3    0.440    0.365  0.125     0.5160          0.2155        0.1140
## 6     2    0.425    0.300  0.095     0.3515          0.1410        0.0775
## 9     3    0.475    0.370  0.125     0.5095          0.2165        0.1125
## 12    3    0.430    0.350  0.110     0.4060          0.1675        0.0810
## 14    1    0.535    0.405  0.145     0.6845          0.2725        0.1710
##    Peso_cascaron Anillos
## 3          0.210       9
## 4          0.155      10
## 6          0.120       8
## 9          0.165       9
## 12         0.135      10
## 14         0.205      10
```

3. [25 puntos] Análisis de Componentes Principales

3(a) [5ptos] Determine las componentes principales de las variables independientes estandarizadas

```
fit.pca <- princomp(abulon[,1:8], cor = TRUE)
fit.pca$loadings
```

```
##
## Loadings:
##                 Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
## Sexo                    0.998
## Longitud        -0.389        -0.156  0.519 -0.171  0.163  0.707
## Diametro        -0.388        -0.143  0.571               -0.701
## Altura          -0.295         0.953
## Peso_total      -0.399        -0.113 -0.293        -0.160         0.845
## Peso_sin_concha -0.388        -0.126 -0.373 -0.607 -0.372        -0.430
## Peso_visceras   -0.387        -0.106 -0.420  0.195  0.765        -0.190
## Peso_cascaron   -0.389                       0.743 -0.468        -0.254
```

```
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
## SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000
## Proportion Var  0.125  0.125  0.125  0.125  0.125  0.125  0.125  0.125
## Cumulative Var  0.125  0.250  0.375  0.500  0.625  0.750  0.875  1.000
```

3(b) [5 ptos] Seleccione las componentes principales que acumulan un 98% de la varianza total. Sugerencia: use `summary()` sobre el modelo de PCA.

```
summary(fit.pca)
```

```
## Importance of components:
##                           Comp.1    Comp.2     Comp.3     Comp.4     Comp.5
## Standard deviation     2.468907 1.0003787 0.71674125 0.42665815 0.31716275
## Proportion of Variance 0.761938 0.1250947 0.06421475 0.02275465 0.01257403
## Cumulative Proportion  0.761938 0.8870327 0.95124746 0.97400210 0.98657613
##                           Comp.6      Comp.7      Comp.8
## Standard deviation     0.27980636 0.151586338 0.078236513
## Proportion of Variance 0.00978645 0.002872302 0.000765119
## Cumulative Proportion  0.99636258 0.999234881 1.000000000
```

Podemos ver que hasta la quinta componente se acumula un 98.7% de la varianza acumulada

3(c) [5 ptos] Interprete las componente principales seleccionadas en (b).

- La primera componente es una ponderación (un overall) de todas las variable excepto el sexo.
- La segunda componente es el sexo.
- La tercera componente es básicamente la altura
- La cuarta componente es un contraste entre las dimensiones (no altura) y los pesos (sin el cascarón).
- La quinta es un contraste entre el peso sin concha y el peso del cascarón.

3(d) [5 ptos] Encuentre las estimaciones de máxima verosimilitud de L y Ψ para $m = 3$ factores de las variables independientes estandarizadas (sin rotación). Nota: Recuerde que L son los *loadings* y Ψ es también conocido como *uniquenesses*.

```
fit <- factanal(abulon[,1:8],3,rotation = "none")

(L = fit$loadings)
```

```
##
## Loadings:
##                Factor1 Factor2 Factor3
## Sexo
## Longitud         0.959   0.205
## Diametro         0.961   0.266
## Altura           0.660           0.107
## Peso_total       0.989
## Peso_sin_concha  0.976  -0.187
```

```
## Peso_visceras    0.943          0.174
## Peso_cascaron    0.947          0.229
##
##                Factor1 Factor2 Factor3
## SS loadings      5.995   0.169   0.116
## Proportion Var   0.749   0.021   0.014
## Cumulative Var   0.749   0.770   0.785
```

```
(Psi = fit$uniquenesses)
```

```
##           Sexo       Longitud       Diametro         Altura     Peso_total
##     0.98969325     0.03878970     0.00500000     0.55256970     0.00500000
## Peso_sin_concha Peso_visceras Peso_cascaron
##     0.00500000     0.07491613     0.05014806
```

3(e) [5 ptos] Interprete los tres factores anteriormente hallados. ¿Qué representa cada factor?

- El primer factor es una ponderación de todas las variables de peso y medidas, sin tener en cuenta el sexo.
- El segundo factor es un contraste entre las dimensiones (sin altura) y el peso sin concha.
- El tercer factor mide la altura el peso de las visceras y el peso del cascaron.

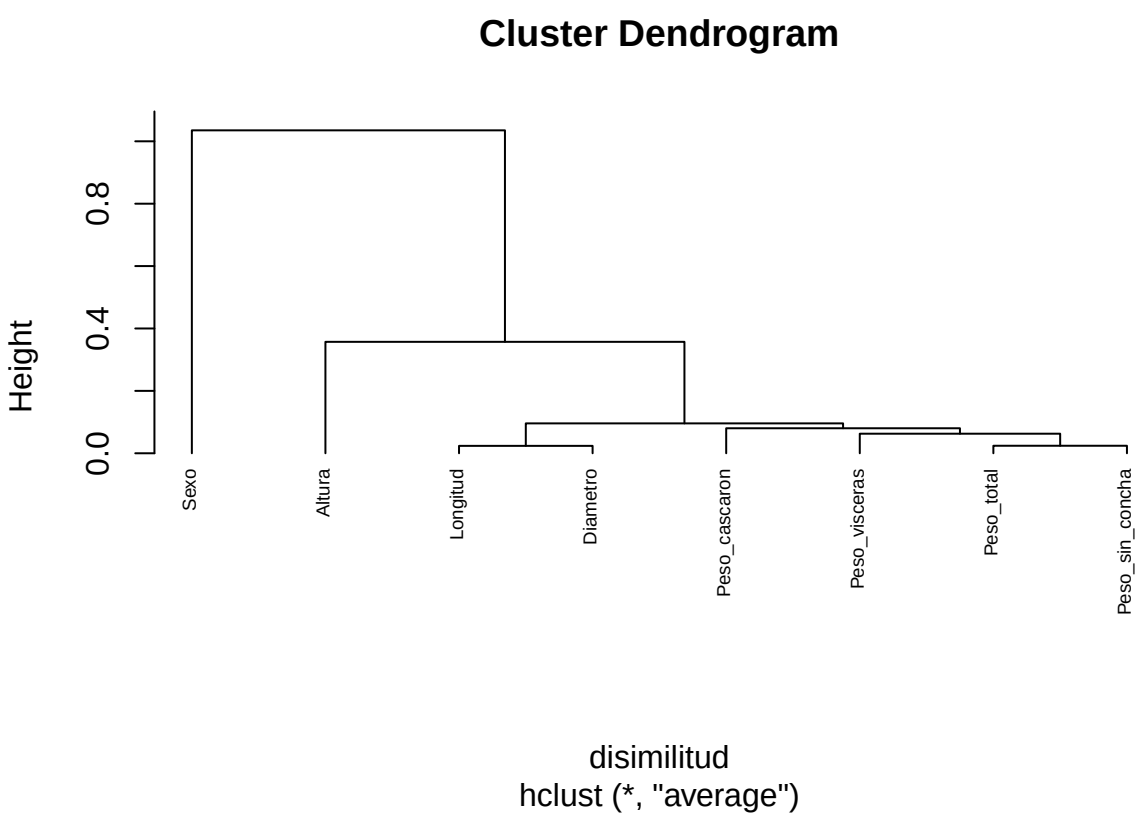
4. [25 ptos] Clustering y Clasificación

4(a) [10 ptos] Realice un clustering jerárquico, no sobre las observaciones sino sobre las variables dependientes, usando la medida de disimilitud como $1 - correlación(X)$ y un vínculo promedio (*average linkage*). Grafique el dendograma.

Use la siguiente medida de disimilitud:

```
disimilitud=as.dist(1-cor(abulon[,1:8]))
```

```
fit.hc <- hclust(disimilitud ,method = "average")
plot(fit.hc, hang = -1, cex = 0.6)
```



4(b) [5 ptos] Observando el dendograma, ¿cuántos clusters usaría para representar las variables?. ¿Qué significado le asocia a cada uno?

El dendograma muestra claramente 4 clusters, el primer cluster del Sexo, el segundo de la altura, el tercer de las medidas "axiales" de Longitud y Diámetro" y el cuarto cluster sobre el peso.

4(c) [10 ptos] Realice una clasificación usando discriminantes lineales, estime el error de clasificación mediante la técnica de *leave-one-out cross-validation* (basta con usar el parámetro `CV=TRUE` de `lda`).

```
fit.lda <- lda(abulon[1:8],abulon[,9],CV=TRUE)
( error = sum(fit.lda$class!=abulon$Anillos)/dim(abulon)[1] )
```

```
## [1] 0.5097832
```