# TD/TP 2/2 Morphométrie M2 ASA

Ce TP reprend pour l'essentiel le document "*Les Iris de Fisher ou Comment se familiariser avec le logiciel R*" par A.B. Dufour, J.R. Lobry, D.Chessel, disponible en téléchargement à l'adresse:

http://pbil.univ-lyon1.fr/R/pdf/bs01.pdf

Plusieurs parties proviennent également du livre de Julien Claude "*Morphometrics with R*" Ed Springer, 2008.

Les données utilisées ici sont célèbres. Elles ont été collectées par Edgar Anderson. Ce sont les mesures en centimètres des variables suivantes : longueur du sépale (Sepal.Length), largeur du sépale (Sepal.Width), longueur du pétale (Petal.Length) et largeur du pétale (Petal.Width) pour trois espèces d'iris : Iris setosa, I. versicolor et I. virginica.

Sir R.A. Fisher a utilisé ces données pour construire des combinaisons linéaires des variables permettant de séparer au mieux les trois espèces d'iris.



On lit le data frame iris.

data(iris)

names(iris)

[1] "Sepal.Length" "Sepal.Width" "Petal.Length" "Petal.Width" "Species"

dim(iris)

[1] 150 5

Voir à quoi correspond les fonction names et dim.

Approche bivariée : on réalise les analyses de la variance à un facteur, variable par variable. On prend par exemple la recherche d'une relation entre la longueur des sépales et les espèces.

Calcul des moyennes par groupe

tapply(iris$Sepal.Length, iris$Species, mean)

setosa versicolor virginica

5.006 5.936 6.588

Calcul des écarts-types par groupe

tapply(iris$Sepal.Length, iris$Species, sd)

setosa versicolor virginica

0.3524897 0.5161711 0.6358796

Analyse de la variance à un facteur

options(show.signif.stars = FALSE)

anova(lm(iris$Sepal.Length ~ iris$Species))

Analysis of Variance Table

Response: iris$Sepal.Length

Df Sum Sq Mean Sq F value Pr(>F)

iris$Species 2 63.212 31.606 119.26 < 2.2e-16

Residuals 147 38.956 0.265

Charger la librairie ade4 et taper le code ci-dessous. Analyser la fonction ainsi créée.

library(ade4)

par(mar = c(0, 0, 0, 0))

pan1 <- function(x, y, ...) {

xy <- cbind.data.frame(x, y)

s.class(xy, iris$Species, include.ori = F, add.p = T, clab = 1.5,

col = c("blue", "black", "red"), cpoi = 2, csta = 0.5)

}

pairs(iris[, 1:4], panel = pan1)



Approche en dimension 3 : on peut encore visualiser en trois dimensions.

library(scatterplot3d)

par(mfrow = c(2, 2))

mar0 = c(2, 3, 2, 3)

scatterplot3d(iris[, 1], iris[, 2], iris[, 3], mar = mar0, color =

c("blue", "black", "red")[iris$Species], pch = 19)

scatterplot3d(iris[, 2], iris[, 3], iris[, 4], mar = mar0, color =

c("blue", "black", "red")[iris$Species], pch = 19)

scatterplot3d(iris[, 3], iris[, 4], iris[, 1], mar = mar0, color =

c("blue", "black", "red")[iris$Species], pch = 19)

scatterplot3d(iris[, 4], iris[, 1], iris[, 2], mar = mar0, color =

c("blue","black", "red")[iris$Species], pch = 19)



Utiliser la fonction lda après avoir consulté la documentation dans le package "MASS"

lda1 <- lda(as.matrix(iris[, 1:4]), iris$Species)

lda1

Call:

lda(as.matrix(iris[, 1:4]), grouping = iris$Species)

Prior probabilities of groups:

setosa versicolor virginica

0.3333333 0.3333333 0.3333333

Group means:

Sepal.Length Sepal.Width Petal.Length Petal.Width

setosa 5.006 3.428 1.462 0.246

versicolor 5.936 2.770 4.260 1.326

virginica 6.588 2.974 5.552 2.026

Coefficients of linear discriminants:

LD1 LD2

Sepal.Length 0.8293776 0.02410215

Sepal.Width 1.5344731 2.16452123

Petal.Length -2.2012117 -0.93192121

Petal.Width -2.8104603 2.83918785

Proportion of trace:

LD1 LD2

0.9912 0.0088

Explorer l'objet lda1 en tapant. Examiner à quoi correspondent les autres paramètres contenus dans lda1.

summary(lda1)

lda1$means

Calculer une combinaison linéaire des variables de départ avec les coefficients qui sont dans la colonne LD1. Idem pour la colonne LD2.

w1 <- as.vector(as.matrix(iris[, 1:4]) %\*% lda1$scaling[, 1])

w1[1:10]

[1] 5.956693 5.023581 5.384722 4.708094 6.027203 5.596840 5.107511 5.500187 4.455445

[10] 5.237953

w1[141:150]

[1] -8.758193 -7.210666 -7.612586 -8.901126 -8.952466 -7.750110 -7.284671 -7.072847

[9] -7.991252 -6.788261

Plusieurs procédures permettent de tester l'existence d'une réelle différence entre les groupes. L'hypothèse nulle est que chaque groupe est un échantillon aléatoire de la même loi normale multivariée. L'écart entre les moyennes par groupe ne serait alors que la conséquence du hasard. C'est l'ANOVA étendue au cas multivarié.



Cas du test de Wilks. Consulter la documentation et internet pour comprendre le principe général du test.

size <- as.matrix(iris[, 1:4])

spec <- iris[, 5]

m1 <- manova(size ~ spec)

summary(m1, test = "Wilks")

Df Wilks approx F num Df den Df Pr(>F)

spec 2 0.023439 199.15 8 288 < 2.2e-16

Residuals 147

Pour éliminer l'effet de taille, on peut diviser chaque mesure par la moyenne géométrique de l'ensemble des mesures. On obtient alors des descripteurs de forme et non de taille. Le code ci-dessous compare l'analyse discriminante opérée selon deux optiques différentes.

miris<-as.matrix(iris[,1:4])

size<-apply(miris,1,prod)^(1/(dim(miris)[2]))

shapeiris<-miris/size

formlda<-lda(miris, as.factor(iris[,5]))

shapelda<-lda(shapeiris, as.factor(iris[,5]))

proj1<-miris%\*%formlda$scaling

proj2<-shapeiris%\*%shapelda$scaling

layout(matrix(c(1,2),1,2))

plot(proj1,pch=(1:3)[as.factor(iris[,5])],asp=1,cex=0.6,xlab="FD1",ylab="FD2",main="Form FDA")

plot(proj2,pch=(1:3)[as.factor(iris[,5])],asp=1,cex=0.6,xlab="FD1",ylab="FD2",main="Shape FDA")

Le but est maintenant de diviser au hasard le tableau de données en deux parties, la première pour chercher une fonction discriminante, la seconde pour déterminer l'espèce à l'aide de cette fonction. On comparera ensuite le résultat obtenu et les vraies valeurs.

echa <- sample(1:150, 50)

tabref <- iris[echa, 1:4]

espref <- iris[echa, 5]

tabsup <- iris[-echa, 1:4]

espsup <- iris[-echa, 5]

lda2 <- lda(tabref, espref)

lda2

Call:

lda(tabref, espref)

Prior probabilities of groups:

setosa versicolor virginica

0.38 0.38 0.24

Group means:

Sepal.Length Sepal.Width Petal.Length Petal.Width

setosa 4.915789 3.289474 1.457895 0.2473684

versicolor 5.821053 2.657895 4.105263 1.2789474

virginica 6.375000 2.991667 5.308333 2.0583333

Coefficients of linear discriminants:

LD1 LD2

Sepal.Length 1.335947 -0.1300696

Sepal.Width 1.119701 1.7648650

Petal.Length -3.537331 -1.8359582

Petal.Width -1.818110 4.7215422

Proportion of trace:

LD1 LD2

0.9845 0.0155

espestim <- predict(lda2, tabsup)$class

table(espestim, espsup)

espsup

espestim setosa versicolor virginica

setosa 31 0 0

versicolor 0 29 0

virginica 0 2 38

A l'aide de 50 plantes, on récupère pratiquement sans erreur le nom des 100 inconnues. Le cas est très favorable.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Exemple des crânes égyptiens**

On étudie ici un cas qui est beaucoup moins favorable avec les données *skulls* proposées par B. Manly [1994]. Les mesures concernent 5 groupes de 30 crânes égyptiens. Les classes sont :

1. période prédynastique ancienne (4000 avant JC) ;
2. période prédynastique récente (3300 avant JC)
3. 12 et 13 ème dynastie (1850 avant JC)
4. période de Ptolémée (200 avant JC)
5. période romaine (150 après JC)

Les variables sont définies comme suit :



Réaliser les mêmes étapes avec le fichier *skulls* contenu dans le package "ade4" après avoir préparé les données (cf ci dessous).

data(skulls)

fac <- gl(5, 30)

levels(fac) <- c("-4000", "-3300", "-1850", "-200", "+150")