Lab 8 Discriminant Analysis (DA)

The goal of this lab is to become familiar with **Discriminant Analysis (DA)**. DA allows you to determine what variables are important in separating groups. It is an eigenanalysis that reduces dimensionality of multivariate data in to new “canonical axes”.

# Set up R session

## Data

Today you will be using the iris data set found in the package datasets in R. To access this data set, simply type:

iris

To learn more about the data set:

?iris

## Install and load packages

We will be using the following packages:

library(dplyr)  
library(MASS)  
library(candisc)  
library(ade4)  
library(vegan)  
library(ggplot2)

# Testing the assumptions of DA

## Homogeneity of within-group variance-covariance matrices

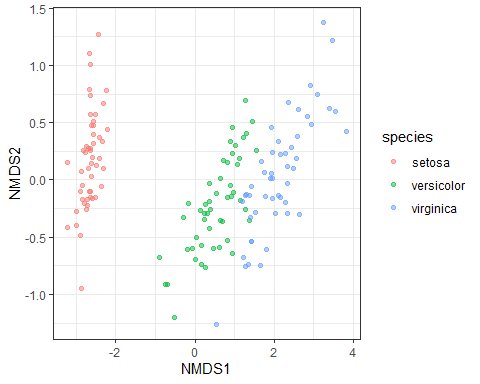
First, lets look at these groups in NMDS space.

dis <- vegdist(iris[,1:4], "euclidean")  
  
iris\_nmds<-metaMDS(dis)

Now lets plot but using ggplot2 this time.

First build a data frame with the first two axes and the groups (i.e., species names):

NMDS=data.frame(NMDS1=iris\_nmds$point[,1],NMDS2=iris\_nmds$point[,2],species=iris$Species)  
  
ggplot(NMDS,aes(x=NMDS1,y=NMDS2))+  
 geom\_point(data=NMDS,aes(x=NMDS1,y=NMDS2, color=species), alpha = 0.5)+  
 viridis::scale\_fill\_viridis() +  
 theme\_bw()



Next, we will use the Fligner-Killeen test of homogeneity of variances. This test has been shown to be robust to departures from normality (which most of our data does to some extent)

?fligner.test

You want to test the variance of each variable (n=4) across all three groups (i.e.’ species). **Remember, you don’t want there to be significant differences**

fligner.test(iris$Sepal.Length, iris$Species)  
fligner.test(iris$Sepal.Width, iris$Species)  
fligner.test(iris$Petal.Length, iris$Species)  
fligner.test(iris$Petal.Width, iris$Species)

We will also use a Multivariate test of homogeneity of variance called ‘betadisper’ in the vegan package.

?betadisper

Calculate multivariate dispersions:

# Euclidean distances between samples  
dis <- vegdist(iris[,1:4], "euclidean")  
  
# groups are the three different species  
groups <- iris$Species  
  
#multivariate dispersions  
MVdisp <- betadisper(dis, groups)  
  
  
#Perform parametric test  
disp\_aov<-anova(MVdisp)  
  
# Tukey's Honest Significant Differences  
MVdisp.HSD <- TukeyHSD(MVdisp)  
MVdisp.HSD

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = distances ~ group, data = df)  
##   
## $group  
## diff lwr upr p adj  
## versicolor-setosa 0.224334 0.05012946 0.3985385 0.0076307  
## virginica-setosa 0.334722 0.16051751 0.5089266 0.0000332  
## virginica-versicolor 0.110388 -0.06381649 0.2845926 0.2937401

## non-parametric test: Permutation test for F  
perm\_MVdisp <- permutest(MVdisp, permutations = 99, pairwise = TRUE)  
perm\_MVdisp

##   
## Permutation test for homogeneity of multivariate dispersions  
## Permutation: free  
## Number of permutations: 99  
##   
## Response: Distances  
## Df Sum Sq Mean Sq F N.Perm Pr(>F)   
## Groups 2 2.9092 1.45458 10.748 99 0.01 \*\*  
## Residuals 147 19.8941 0.13533   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Pairwise comparisons:  
## (Observed p-value below diagonal, permuted p-value above diagonal)  
## setosa versicolor virginica  
## setosa 1.0000e-02 0.01  
## versicolor 4.5862e-04 0.25  
## virginica 2.6235e-05 1.7926e-01

Should you transform the iris variables?

If you think yes, use the code below. If you think no, skip ahead.

log<-cbind.data.frame(apply(iris[,1:4]+1,2,log),iris$Species)  
names(log)[5]<-"Species"

Re-run the Fligner-Killeen test on the transformed data:

fligner.test(log$Sepal.Length,log$Species)  
fligner.test(log$Sepal.Width ,log$Species)  
fligner.test(log$Petal.Length ,log$Species)  
fligner.test(log$Petal.Width, log$Species)

Re-run multivariate homogeneity of variance test

# Euclidean distances between samples  
log\_dis <- vegdist(log[,1:4], "euclidean")  
  
  
# groups are the three different species  
groups <- iris$Species  
  
## Calculate multivariate dispersions  
log\_MVdisp <- betadisper(log\_dis, groups)  
  
  
#Perform parametric test  
disp\_aov<-anova(log\_MVdisp)  
  
  
# Tukey's Honest Significant Differences  
MVdisp.HSD <- TukeyHSD(log\_MVdisp)  
MVdisp.HSD

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = distances ~ group, data = df)  
##   
## $group  
## diff lwr upr p adj  
## versicolor-setosa 0.017189486 -0.01686398 0.05124295 0.4579384  
## virginica-setosa 0.020275571 -0.01377789 0.05432903 0.3385163  
## virginica-versicolor 0.003086085 -0.03096738 0.03713955 0.9749425

# non-parametric test: Permutation test for F  
perm\_MVdisp <- permutest(log\_MVdisp, permutations = 99, pairwise = TRUE)  
perm\_MVdisp

##   
## Permutation test for homogeneity of multivariate dispersions  
## Permutation: free  
## Number of permutations: 99  
##   
## Response: Distances  
## Df Sum Sq Mean Sq F N.Perm Pr(>F)  
## Groups 2 0.01194 0.0059675 1.1539 99 0.29  
## Residuals 147 0.76020 0.0051714   
##   
## Pairwise comparisons:  
## (Observed p-value below diagonal, permuted p-value above diagonal)  
## setosa versicolor virginica  
## setosa 0.27000 0.18  
## versicolor 0.23860 0.90  
## virginica 0.14779 0.83459

## Multivariate Normality

To look for multivariate normality, let’s look at the distribution of each variable for each group and run (overly conservative) mulitvariate normality tests for both the log and untransformed Iris data.

#filter for each species: e.g., setosa below  
  
setosa<-dplyr::filter(iris, Species == "setosa")  
  
#untransformed  
  
mshapiro.test(t(setosa[,1:4]))  
  
mvn(setosa[,1:4], mvnTest = "mardia")  
  
#log  
  
mshapiro.test(t(log(setosa[,1:4])))  
  
mvn(log(setosa[,1:4]), mvnTest = "mardia")

## Multicolinearity

To test for multicolinearity, you should look at all pairwise correlations. Remember, correlations > 0.7 can be trouble.

cor(iris[,1:4])

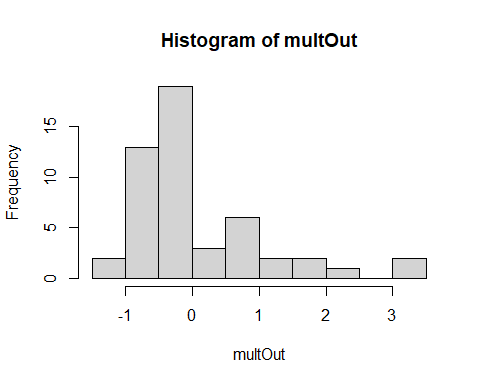
## Sepal.Length Sepal.Width Petal.Length Petal.Width  
## Sepal.Length 1.0000000 -0.1175698 0.8717538 0.8179411  
## Sepal.Width -0.1175698 1.0000000 -0.4284401 -0.3661259  
## Petal.Length 0.8717538 -0.4284401 1.0000000 0.9628654  
## Petal.Width 0.8179411 -0.3661259 0.9628654 1.0000000

We have some co-linearity. Let’s move ahead using all four variables, but we will also run the analysis without Petal.Length.

## Outliers

Outliers can have a large influence on canonical axes in DA. You are going to take a multivariate approach to identifying outliers for both the untransformed and transformed data (we did this in lab 2):

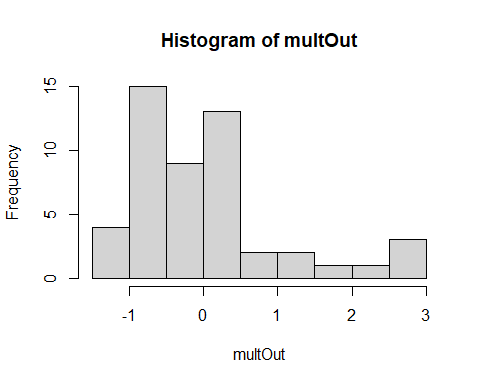
#Calculate a withi-group distance matrix:  
#setosa  
  
eucDist<- vegdist(iris[1:50,1:4],"euclidean")  
  
#Calculate the average distance of each sample to all other samples (i.e. column average) and turn the means in z-scores:  
  
multOut<-scale(colMeans(as.matrix(eucDist)))  
  
  
#Now look at a histogram of these data to identify samples that are > 3 sd from the mean:  
  
hist(multOut)



#and get the number of those samples:  
  
setosa\_out<-multOut [multOut >3,]  
setosa\_out

## 16 42   
## 3.044307 3.294217

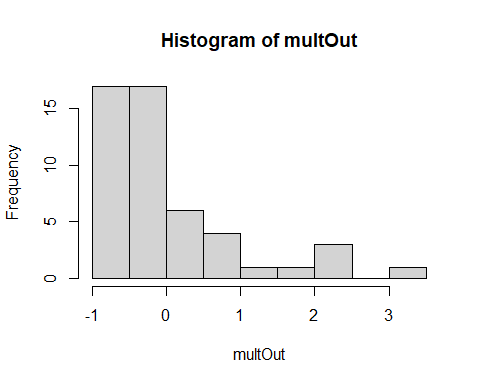
#Repeat for the other two groups (i.e.,species):  
  
#versicolor  
eucDist<- vegdist(iris[51:100,1:4],"euclidean")  
multOut<-scale(colMeans(as.matrix(eucDist)))  
hist(multOut)



versicolor\_out<-multOut [multOut >3,]  
versicolor\_out

## numeric(0)

#virginica  
  
eucDist<- vegdist(iris[101:150,1:4],"euclidean")  
multOut<-scale(colMeans(as.matrix(eucDist)))  
hist(multOut)



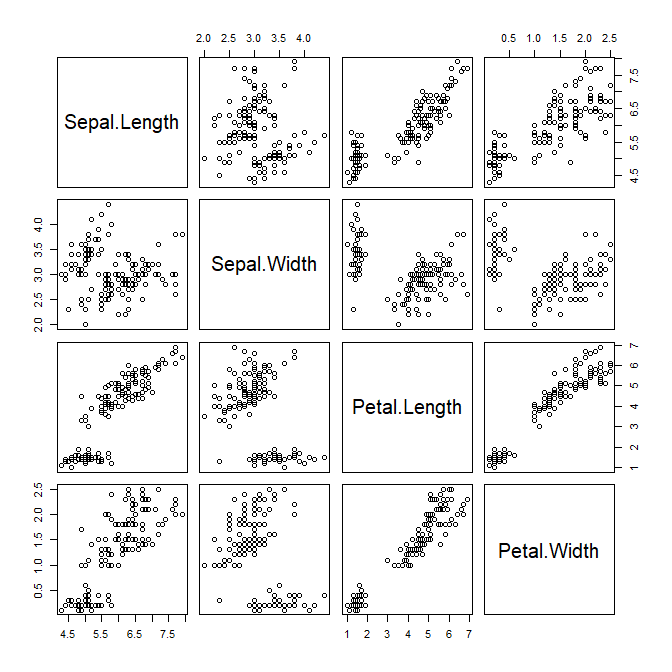
virginica\_out<-multOut [multOut >3,]  
  
#Finally, make a vector of the outliers to pull out of the data set later:  
  
Outliers<-c(setosa\_out,versicolor\_out,virginica\_out)  
Outliers

## 16 42 107   
## 3.044307 3.294217 3.137975

## Linearity

Next we test for linear relationship between variables. This is key to making sure that the variables change in a linear fashion along underlying gradients (i.e. canonical axes):

pairs(iris[,1:4])



Believe it or not, those pass as linear relationships!

# Discriminant analysis

You will use the LDA function in the *MASS* package to conduct DA and will also use the candisc function in package *candisc*.

?lda  
  
?candisc

First run the analysis on the un-transformed data with all of the variables. We are going to split our data set into “training” data and “testing” data. We randomly select 75 samples from the iris data set:

set.seed(11)  
train <- sample(1:150, 75)

Check the frequency of each species in the training data, which will help set your priors:

prior<-table(iris$Sp[train])

Next run the DA for the training data:

iris.LDA <- lda(Species ~ ., iris, prior = cbind(prior/75), subset = train)

**“Species ~ .” is the model and the “dot” stands for all of the variables (so you don’t have to type them all)**

What do the results tell you about the number of meaningful canonical axes and the absolute contribution of each variable to those axes?

# Assessing and interpreting canonical axes

## Relative % Criterion

The “proportion of trace” from the lda output is the **Relative % criterion**.

## Canonical Correlation Criterion

Here, you want to test the correlation between the canonical scores and the grouping variable. First you need to predict the membership (i.e. calculate canonical scores) of each sample in the training set. You will use the predict.lda function from the *MASS* package:

?predict.lda

iris.LDA.p<-predict(iris.LDA, iris[train, ])

You want to test the canonical scores (iris.LDA.P$x) with the grouping variable (i.e. species):

corTest<-lm(iris.LDA.p$x~iris$Sp[train])  
summary(corTest)

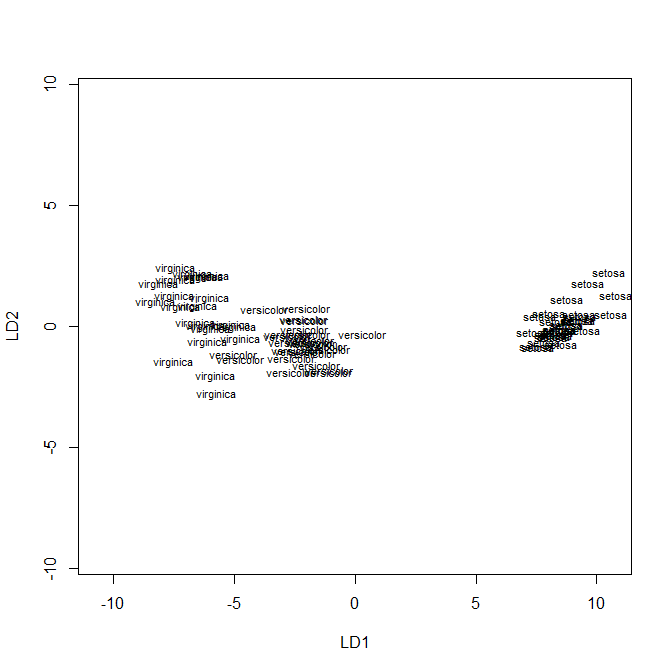
## Response LD1 :  
##   
## Call:  
## lm(formula = LD1 ~ iris$Sp[train])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.82910 -0.56549 0.02994 0.52689 2.53286   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.1061 0.1890 42.89 <2e-16 \*\*\*  
## iris$Sp[train]versicolor -10.9323 0.2752 -39.73 <2e-16 \*\*\*  
## iris$Sp[train]virginica -15.2115 0.2849 -53.39 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1 on 72 degrees of freedom  
## Multiple R-squared: 0.9777, Adjusted R-squared: 0.977   
## F-statistic: 1575 on 2 and 72 DF, p-value: < 2.2e-16  
##   
##   
## Response LD2 :  
##   
## Call:  
## lm(formula = LD2 ~ iris$Sp[train])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.3421 -0.5492 -0.0659 0.4359 2.1088   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.1742 0.1890 0.922 0.35959   
## iris$Sp[train]versicolor -0.8680 0.2752 -3.154 0.00235 \*\*  
## iris$Sp[train]virginica 0.3923 0.2849 1.377 0.17278   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1 on 72 degrees of freedom  
## Multiple R-squared: 0.2169, Adjusted R-squared: 0.1952   
## F-statistic: 9.972 on 2 and 72 DF, p-value: 0.0001504

How does the correlation and significance look for each canonical axis?

## Plot canonical scores and axes

Now visually asses the discrimination of the three species of iris by the canonical axes:

plot(iris.LDA, xlim=c(-11,11), ylim=c(-6,6))



## Classification accuracy

You now want to measure how well the axes discriminate. The higher the correct classification rate, the greater degree of group discrimination achieved by the canonical axes. We will classify the training data first. Create a classification matrix:

ct<- table( iris[train, ]$Species, predict(iris.LDA, iris[train, ])$class)   
  
#Change to a table of proportions:  
  
pct<-prop.table(ct)  
pct

##   
## setosa versicolor virginica  
## setosa 0.37333333 0.00000000 0.00000000  
## versicolor 0.00000000 0.30666667 0.02666667  
## virginica 0.00000000 0.00000000 0.29333333

#Calculate classification rate by summing the diagonal:  
  
sum(diag(pct))

## [1] 0.9733333

## Interpreting canonical axes (raw coefficients, standardized weights, and structure coefficients)

Unfortunately, lda doesn’t provide an easy way to calculate the standardized weights, and structure coefficients. For this, you will use candisc in the *candisc* package.

?candisc

For candisc, you build a linear model using lm:

iris.mod <- lm(cbind(Sepal.Length, Sepal.Width, Petal.Length, Petal.Width) ~ Species, data=iris[train, ])

And run the candisc function:

iris.can <- candisc(iris.mod, data=iris[train, ])

Next pull the raw coefficients, standardized weights, and structure coefficients from the output:

iris.can$coeffs.raw

## Can1 Can2  
## Sepal.Length 0.1720065 -0.4097189  
## Sepal.Width 2.5056884 -1.3726830  
## Petal.Length -2.1445915 1.5937114  
## Petal.Width -3.0285303 -3.7825246

iris.can$coeffs.std

## Can1 Can2  
## Sepal.Length 0.08134361 -0.1937602  
## Sepal.Width 0.85633837 -0.4691250  
## Petal.Length -0.82117521 0.6102404  
## Petal.Width -0.64583534 -0.8066249

iris.can$structure

## Can1 Can2  
## Sepal.Length -0.7882513 -0.26012113  
## Sepal.Width 0.6131572 -0.63022074  
## Petal.Length -0.9871812 -0.04062721  
## Petal.Width -0.9673749 -0.25238724

Note that the raw coefficients are the same as the lda output:

iris.LDA$scaling

## LD1 LD2  
## Sepal.Length 0.1720065 0.4097189  
## Sepal.Width 2.5056884 1.3726830  
## Petal.Length -2.1445915 -1.5937114  
## Petal.Width -3.0285303 3.7825246

*notice that the signs can sometimes be flipped but coefficients are the same*

Looking at the structure coefficients, how would you “define” the first canonical axis?

# Validating canonical axes

A DA is only as good as its ability to classify new data correctly. Here, you are going to use the “testing” data from your split sample to see how well the canonical axes predict group membership of “new” samples. You are going to use the predict.lda function again. This time with the testing data (i.e., iris[-train, ])

iris.LDA.new<-predict(iris.LDA , iris[-train, ])

Next, make a classification table and calculate the classification rate:

ynew.table<-table(iris[-train,]$Species,iris.LDA.new$class)  
ynew.table

##   
## setosa versicolor virginica  
## setosa 22 0 0  
## versicolor 0 25 0  
## virginica 0 1 27

sum(diag(prop.table(ynew.table)))

## [1] 0.9866667

How did the canonical axes handle the new data?

# MANOVA, *the other side of the coin of DA*

In MANOVA, we are interested if groups differ in their measured variables. In DA, we are interested in a linear combination of variables that maximize differences between groups. In this case, MANOVA can provide a test of whether or not the groups are different.

? manova

Y<-as.matrix(iris[,1:4])  
Sp<-factor(iris[,5])  
  
fit <- manova(Y ~ Sp)  
summary(fit,test="Wilks")

## Df Wilks approx F num Df den Df Pr(>F)   
## Sp 2 0.023439 199.15 8 288 < 2.2e-16 \*\*\*  
## Residuals 147   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Post-hoc tests:

Yset<-as.matrix(iris[1:50,1:4])  
Yversi<-as.matrix(iris[51:100,1:4])  
Yvirg<-as.matrix(iris[101:150,1:4])  
Sp<-factor(iris[,5])  
  
fit1 <- manova(rbind(Yset,Yversi) ~ Sp[1:100])  
summary(fit1,test="Hotelling-Lawley")

## Df Hotelling-Lawley approx F num Df den Df Pr(>F)   
## Sp[1:100] 1 26.335 625.46 4 95 < 2.2e-16 \*\*\*  
## Residuals 98   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

fit2 <- manova(rbind(Yversi,Yvirg) ~ Sp[51:150])  
summary(fit2,test="Hotelling-Lawley")

## Df Hotelling-Lawley approx F num Df den Df Pr(>F)   
## Sp[51:150] 1 3.6273 86.148 4 95 < 2.2e-16 \*\*\*  
## Residuals 98   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

fit3 <- manova(rbind(Yset,Yvirg) ~ Sp[-c(51:100)])  
summary(fit3,test="Hotelling-Lawley")

## Df Hotelling-Lawley approx F num Df den Df Pr(>F)   
## Sp[-c(51:100)] 1 49.792 1182.6 4 95 < 2.2e-16 \*\*\*  
## Residuals 98   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

If you have time, now run the analysis on the log transformed data without the variable “Petal.Length” and the outliers. **Remove the Petal.Length column and outliers before analyzing**.

Does meeting the assumptions of DA change our interpretation of our results?