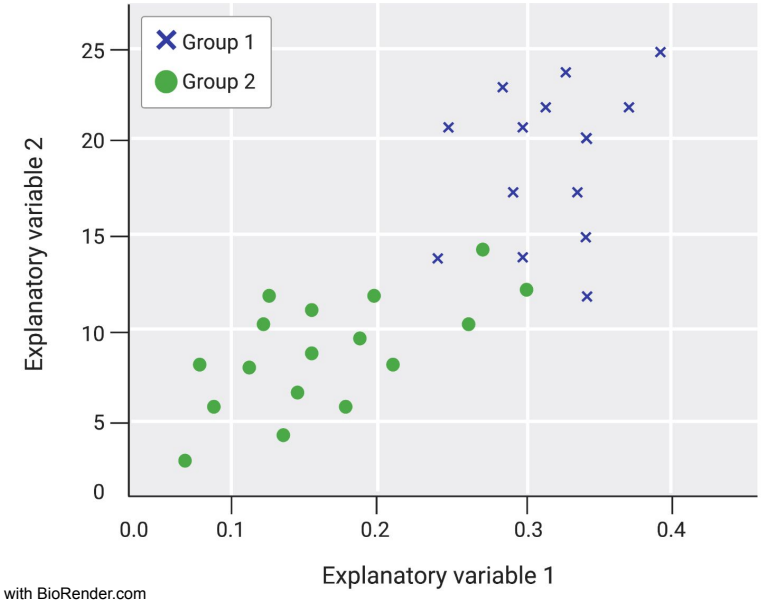
Linear discriminant function analysis

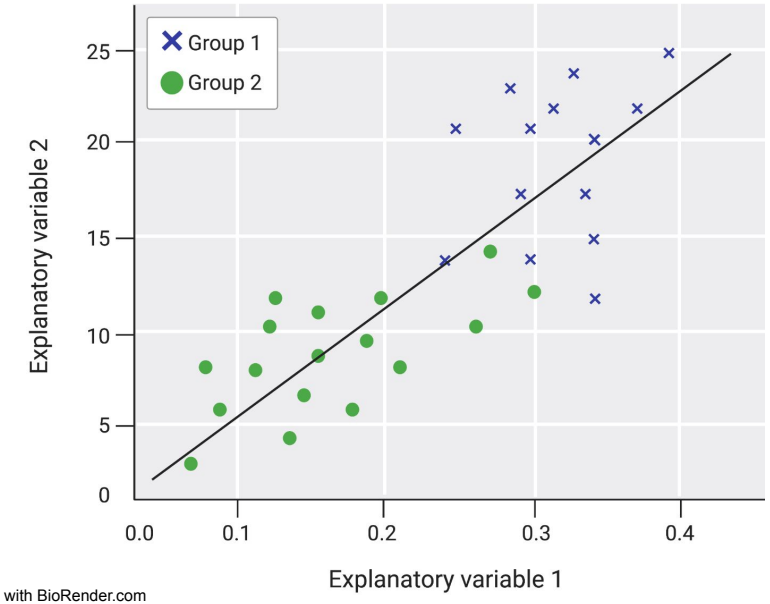
What is Linear Discriminant Analysis?

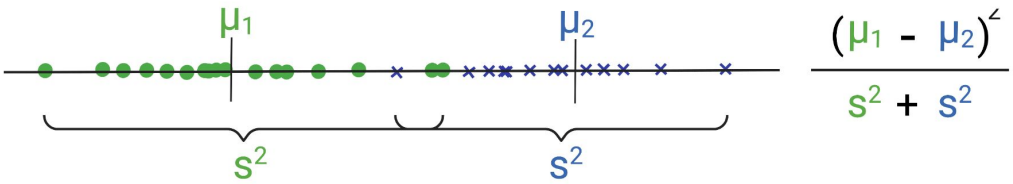
A dimensional reduction technique that uses linear combinations of continuous variables to discriminate between naturally occurring groups. The method simultaneously tries to maximize distance between group means and minimize variation within categories.



What is Linear Discriminant Analysis?

A dimensional reduction technique that uses linear combinations of continuous variables to discriminate between naturally occurring groups. The method simultaneously tries to maximize distance between group means (μ) and minimize variation (s^2) within categories.

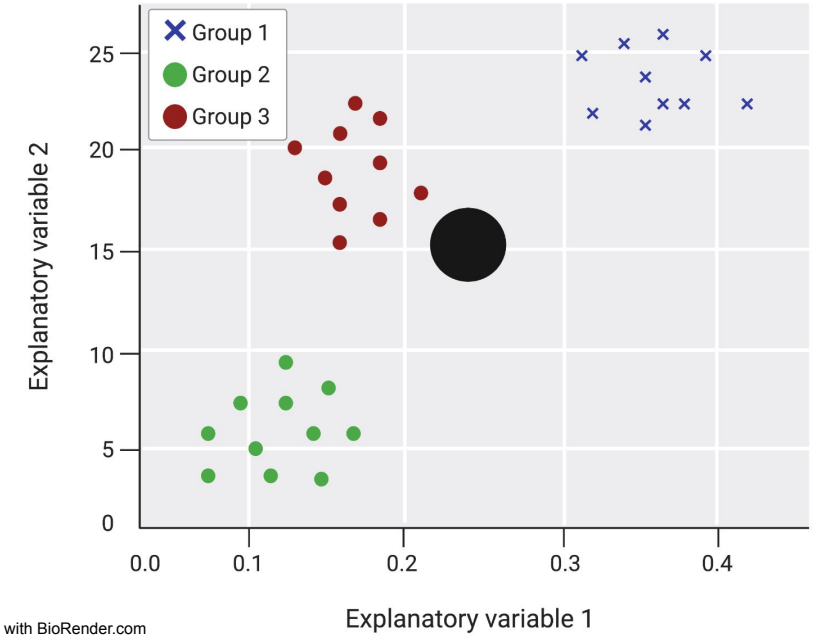




Created with BioRender.com

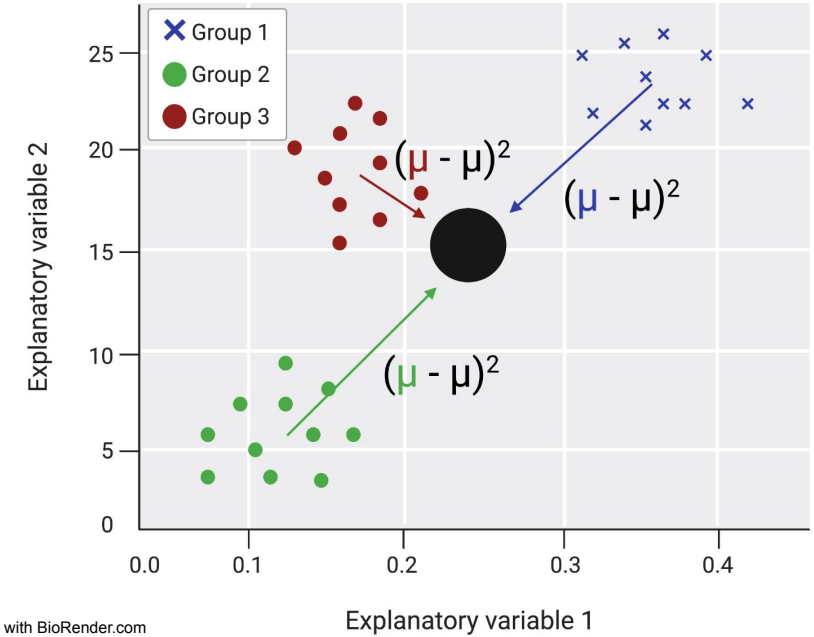
Linear discriminant analysis with more than two categories

The number of discriminant functions output from the analysis is one less than the number of categories. For three groups, two discriminant functions are created.



Linear discriminant analysis with more than two groups

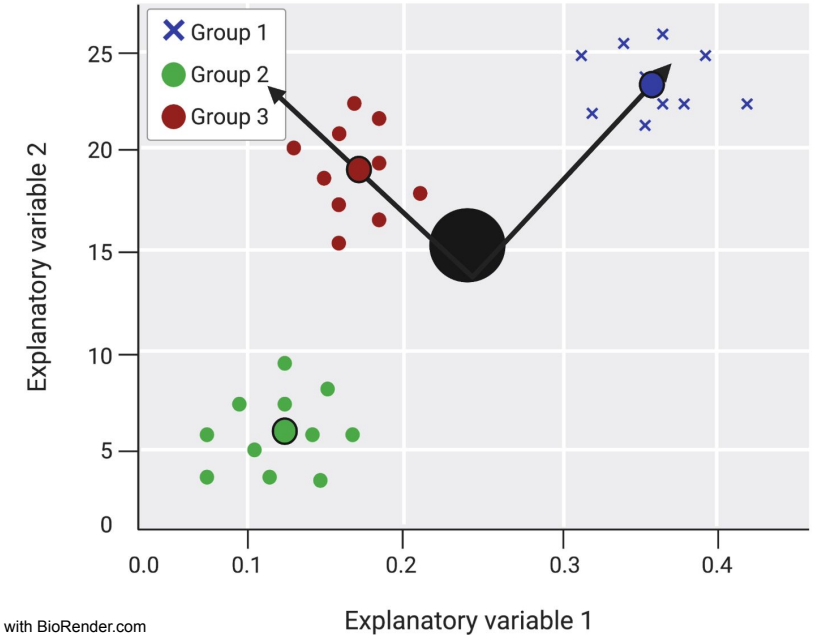
The number of discriminant functions output from the analysis is one less than the number of categories. For three groups, two discriminant functions are created.

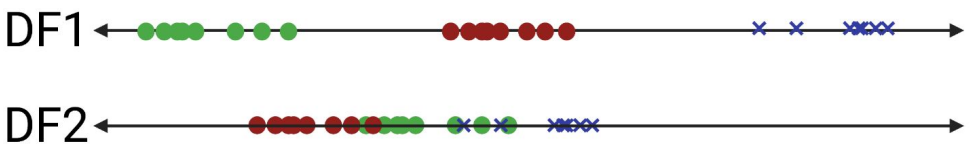


$$\frac{(\mu - \mu)^2 + (\mu - \mu)^2 + (\mu - \mu)^2}{s^2 + s^2 + s^2}$$

Linear discriminant analysis with more than two groups

The number of discriminant functions output from the analysis is one less than the number of categories. For three groups, two discriminant functions are created.





Assumptions

- Each sample must belong to only one group
- Each sample must be independent
- The number of samples in each group must be similar
- Sample size must be at least the number of independent variables + 2.
- Each group must be homoscedastic.
- All independent variables must be multivariate normal.

When should you use it?

LDA is used when you have a dataset with certain pre-defined categories, and you want to understand how well your variables measured can predict the categories that your samples belong in

Example 1: You have taken morphological measurements from many different species. You want to see if the measurements can predict which species your samples belong to

Example 2: You have data on patient lifestyle from individuals that have different chronic illnesses. You want to see whether any of the lifestyle measurements predict disease.

Performing LDA on larval fish behavior data

Larval red drum fish (*Scianops ocellatus*) exhibit stable differences in predator evasion and foraging behavior in the laboratory.

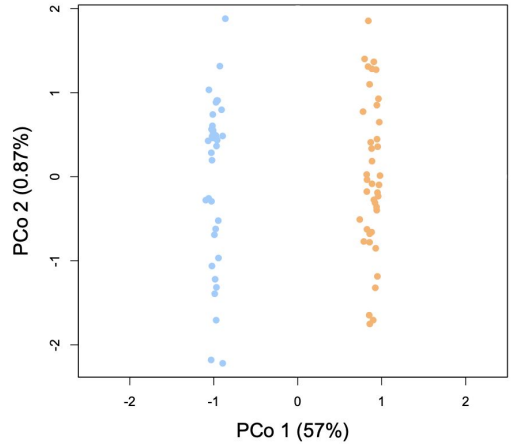
76 larvae were allowed to freely swim in a tank for 5 minutes during which we took measurements which correlate with foraging behavior in this species.

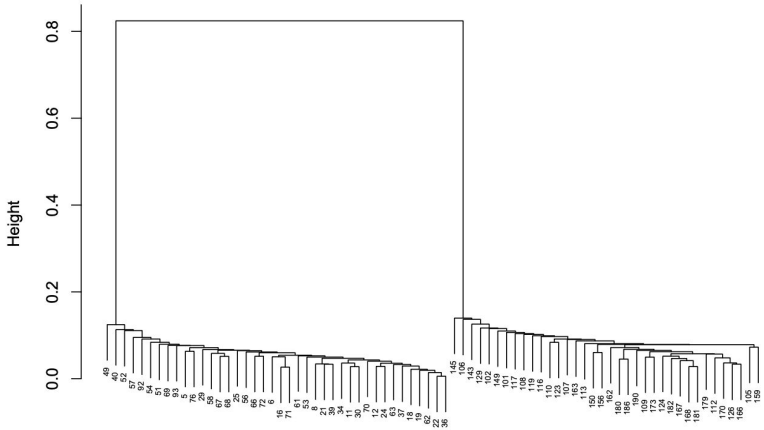
The larvae came from one of two source tanks and genotyping tells us all of the larvae in each tank had the same mother.

Do individuals from different maternal lines behave differently?

Smith et al. (*in review*).

Performing LDA on larval fish behavior data





Larval red drum fish (*Scianops ocellatus*) exhibit stable differences in predator evasion and foraging behavior in the laboratory.

76 larvae were allowed to freely swim in a tank for 5 minutes during which we took measurements which correlate with foraging behavior in this species.

The larvae came from one of two source tanks and genotyping tells us all of the larvae in each tank had the same mother.

Do individuals from different maternal lines behave differently?

Smith et al. (*in review*).

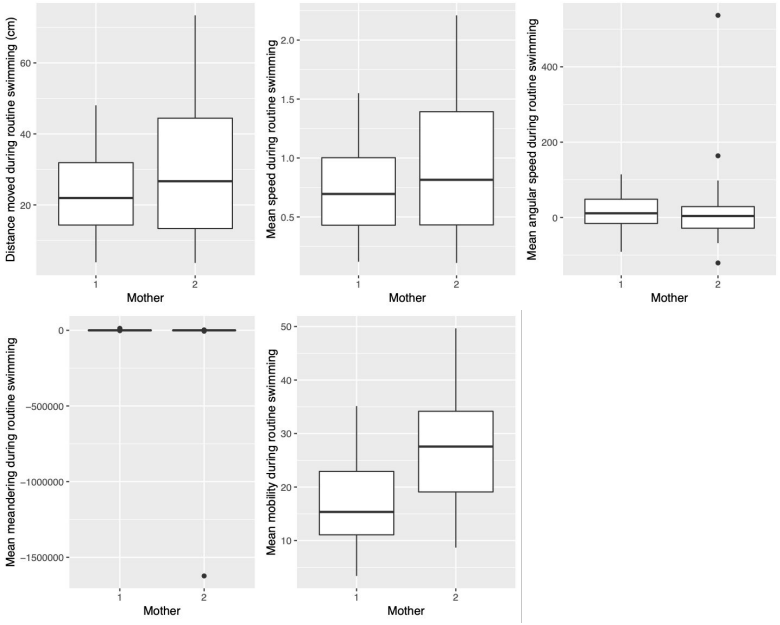
Data processing

	ID	Mother	RDistancedMoved_cm	RMeanSpeed	RMeanAngSpd	RMeanMeander	RMeanMobility
1	101	2	8.98	0.44	536.78	1604.84	18.18
2	102	2	45.99	1.38	16.67	-509.79	25.60
3	105	2	18.70	0.56	-11.91	300.74	14.51
4	106	2	20.53	0.61	-6.29	-204.81	12.97
5	107	2	22.78	0.69	-39.47	-1686.17	26.25
6	108	2	71.05	2.12	52.24	73.21	33.70
7	109	2	43.91	1.32	-68.15	2524.68	30.18
8	110	2	26.63	0.90	-26.10	-352.89	28.88
9	112	2	12.25	0.37	81.17	1033.78	12.78
10	113	2	6.10	0.18	21.73	-2453.87	8.68

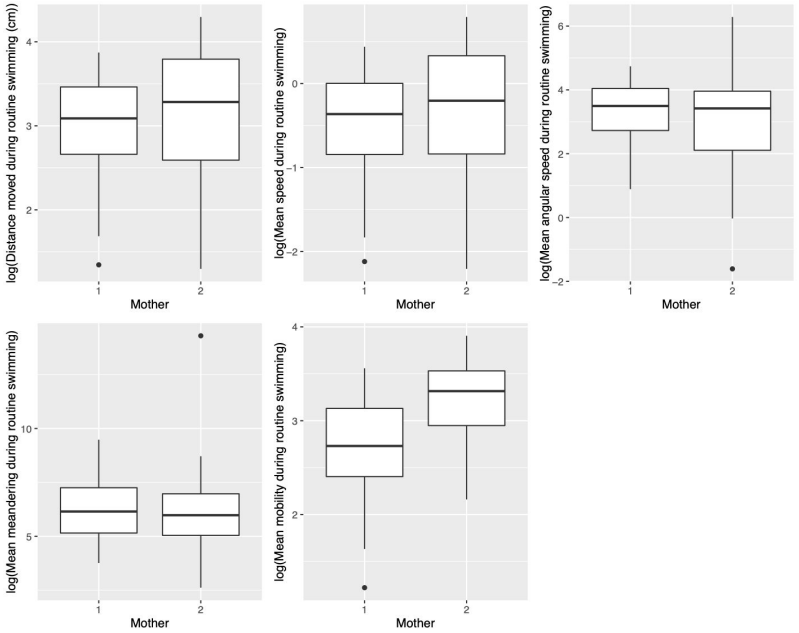
Do we have equal variance in all the variables between groups (maternal id)?

Data processing: equal variance? No.

Raw data



Log transformed data



Data processing and running the model

We split the data into a training dataset (80%) and a test dataset (20%) and normalized the data.

```
library(caret)
set.seed(60)
training.samples <- logtrans_traits$Mother %>% createDataPartition(p = 0.8, list = FALSE)
train.data <- logtrans_traits[training.samples,]
test.data <- logtrans_traits[-training.samples,]
hist.data.frame(train.data,nclass=20)
# normalize the data
preproc.param <- train.data %>% preProcess(method = c("center","scale"))
# transform using estimated parameters
train.transf <- preproc.param %>% predict(train.data)
test.transf <- preproc.param %>% predict(test.data)
```

We run the model using MASS::lda

```
library(MASS)
model <- lda(Mother ~ logSpeed + logDist + logMobility + logAngSpd + logMeander, data=train.transf)
```

Model output

```
Call:
lda(Mother ~ logSpeed + logDist + logMobility + logAngSpd + logMeander,
    data = train.transf)

Prior probabilities of groups:
        1         2
0.5245902 0.4754098

Group means:
    logSpeed    logDist logMobility   logAngSpd  logMeander
1 -0.1479635 -0.1693876  -0.4750878  0.08442790  0.06773704
2  0.1632701  0.1869105   0.5242348 -0.09316182 -0.07474432

Coefficients of linear discriminants:
                   LD1
logSpeed    -5.06929840
logDist      3.86313028
logMobility  1.96716331
logAngSpd    0.19713057
logMeander  -0.01518053
```

Prior probabilities of groups: probability a sample will fall into either group before the data is collected.

Group means: mean values of each sample for each group.

Coefficients of linear discriminants: Coefficients of the resulting LD equation.

Predicting group membership

```
pred_test <- model %>% predict(test.transf)
```

```
$class
 [1] 1 2 1 2 2 2 2 1 1 1 1 2 1 2 1
Levels: 1 2

$posterior                  $x
            1          2                LD1
1  0.92635266 0.07364734   1  -1.36271997
5  0.12393255 0.87606745   5   1.22880097
10 0.66673042 0.33326958   10 -0.30101228
28 0.13241745 0.86758255   28  1.18493918
34 0.36355656 0.63644344   34  0.42279248
35 0.08091516 0.91908484   35  1.50268120
37 0.12678766 0.87321234   37  1.21376319
43 0.91573596 0.08426404   43 -1.27829650
48 0.51106178 0.48893822   48  0.07387311
49 0.81881973 0.18118027   49 -0.77161866
58 0.57372821 0.42627179   58 -0.07212672
59 0.37414295 0.62585705   59  0.39653085
61 0.94490888 0.05509112   61 -1.54181422
66 0.34384268 0.65615732   66  0.47260302
71 0.93266414 0.06733586   71 -1.41838015
```

pred_test$class: predicted group for each sample in the test set.

pred_test$posterior: probability of being assigned to either group.

pred_test$x: LD1 scores for each sample in the test set.

Actual group

Predicted group	Group 1	Group 2
Group 1	**6**	2
Group 2	2	**5**

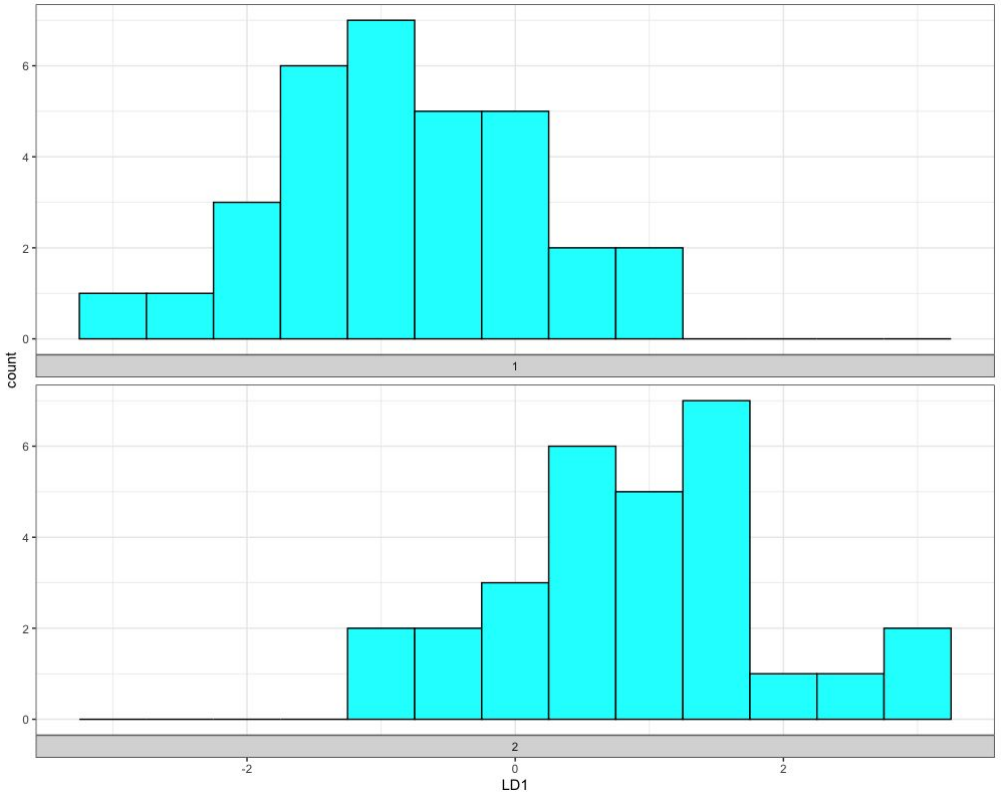
table(test.transf$Mother,pred_test$class)

On diagonal = correct assignment
Off diagonal = incorrect assignment

mean(pred_test$class == test.transf$Mother)

73% accuracy

Visualizing group discrimination



Routine swimming behavior does an okay job of distinguishing between individuals born of different mothers.

This suggests that routine swimming is partially genetically or epigenetically determined.

Actual group

Predicted group	Group 1	Group 2
Group 1	**6**	2
Group 2	2	**5**

73% accuracy

Performing LDA on stress response in ARTs

Swordtails (Xiphophorus nigrensis) males belong to one of three alternative reproductive tactics (ARTs)

Large males court females, small males are coercive, and intermediate males perform both behaviors

They experience many different ecological pressures- large males face more predation, and small males are not preferred over large males by females.

Question: Do their stress responses vary in addition to their mating behaviors?

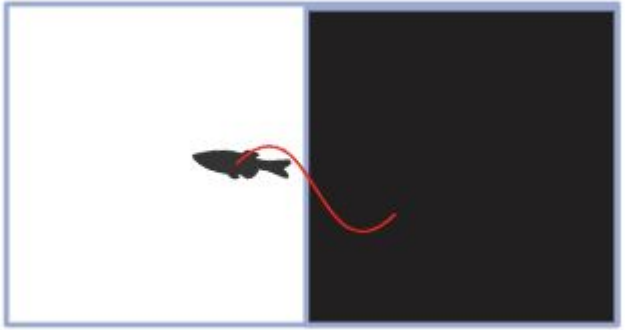
Img: Maderspacher 2010

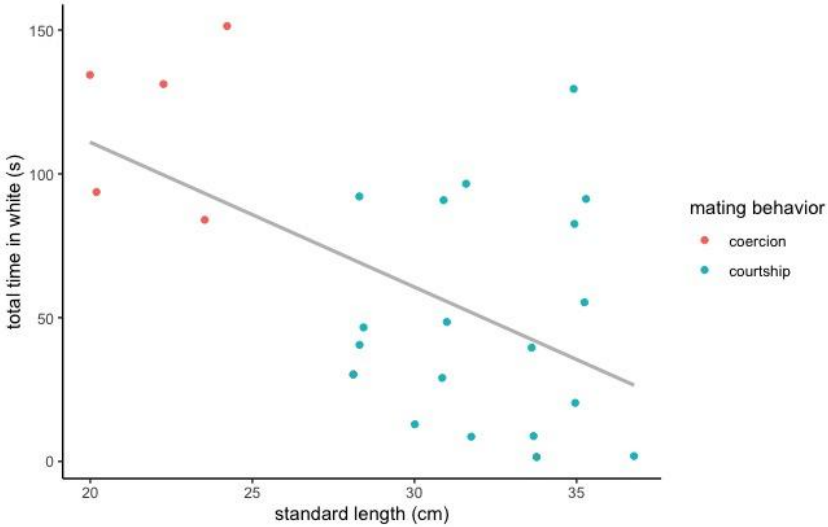
Scototaxis assay

A half white and half dark tank was used to understand stress response behavior.

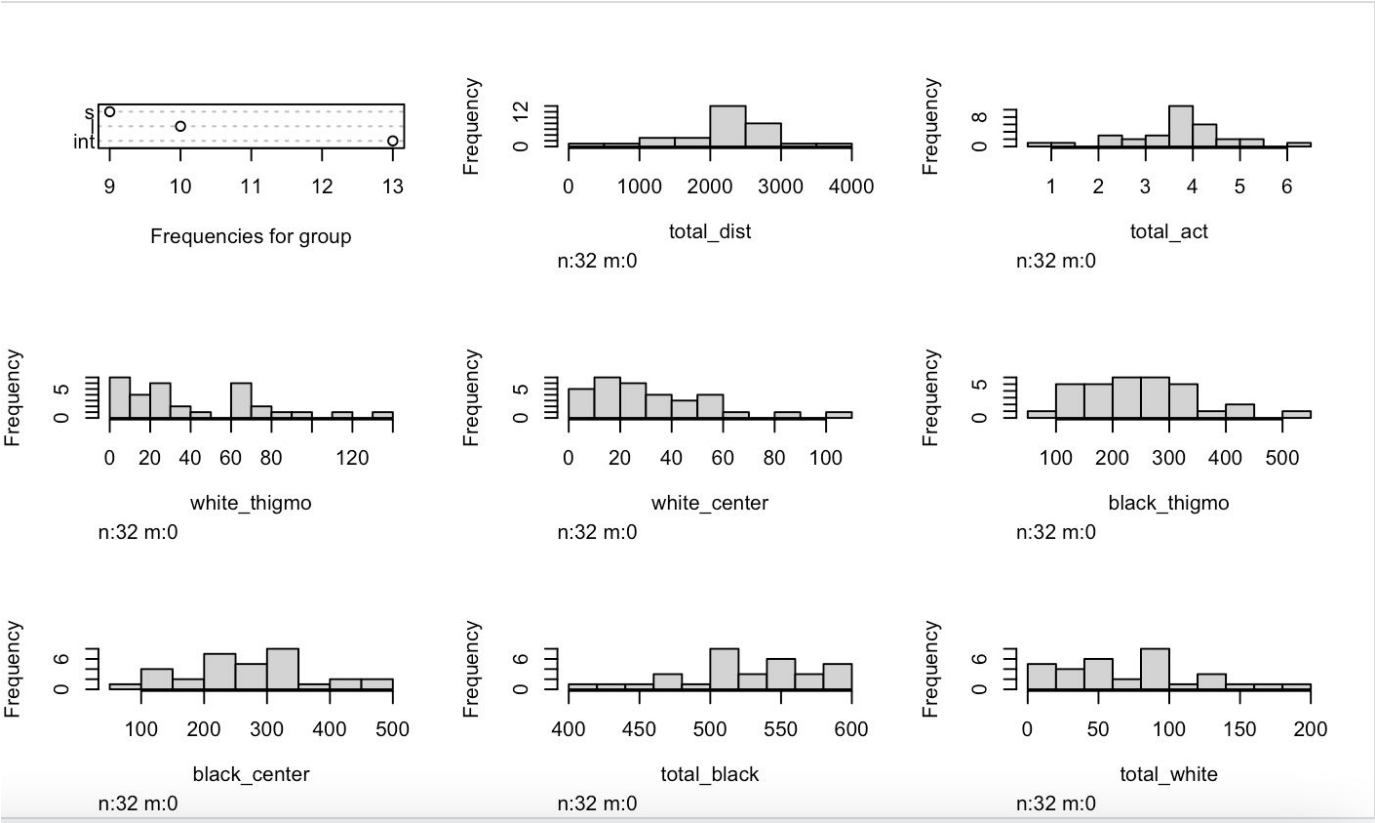
More time spent in the white zones (especially center) indicates boldness, more time in dark zone (especially tank edges) indicates anxiety.

It was shown that time in white zone varies with size class. Thus, we wanted to see if scototaxis metrics could predict size class

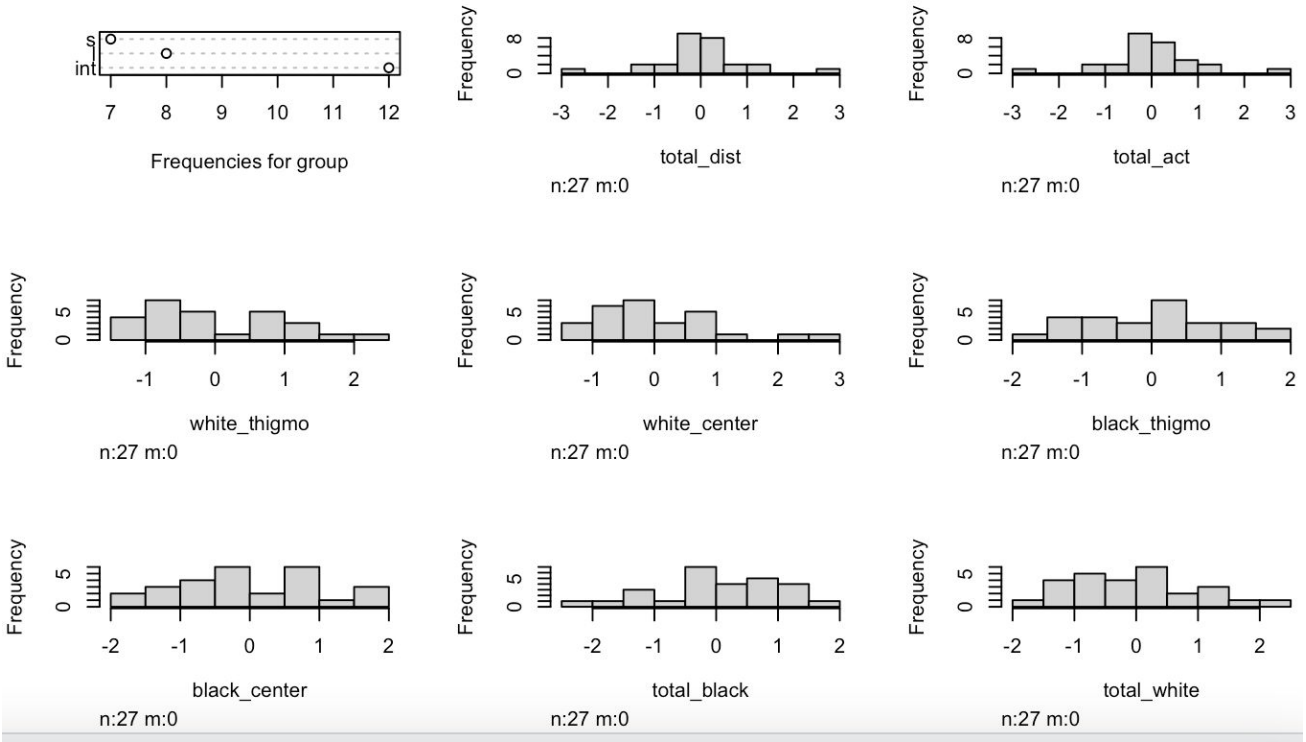




Checking distribution



Preprocessing- centering and scaling the data



MASS::lda() and predict() output

lda = lda(dataframe, group ~ variables)

predictions = predict(lda)

predictions$class shows the classification into groups according to the LDA

predictions$posterior shows the probability of each sample belonging to each group according to the LDA. Here, sample 1 has the most probability of belonging to group “int”

```
> predict(training_lda)
$class
 [1] int int int int int s   int int int int int l   l   int
[15] int int s   int int l   int int s   int int s   s
Levels: int l s

$posterior
          int          l          s
1  0.90377900 0.06798272 0.02823828
2  0.76767458 0.20388099 0.02844443
3  0.49303994 0.44284939 0.06411067
4  0.52705066 0.36632335 0.10662599
5  0.52500820 0.37006388 0.10492792
6  0.21605130 0.21915312 0.56479558
7  0.35582374 0.34104214 0.30313412
```

Validating the LDA: does it predict the groups in dataset?

	group	original_assignments	lda_assignments
	<chr>	<int>	<int>
1	int	9	10
2	l	7	6
3	s	6	6

	int	l	s
int	6	2	1
l	2	4	1
s	2	0	4

Table 1- How many samples assigned to original dataset vs to the groups made by LDA

Table 2- The rows are original data, columns are predictions from LDA, so if the LDA predicted all group assignments, then diagonal would have all cases, and all values not on diagonal = 0. The predictive accuracy is 64%.

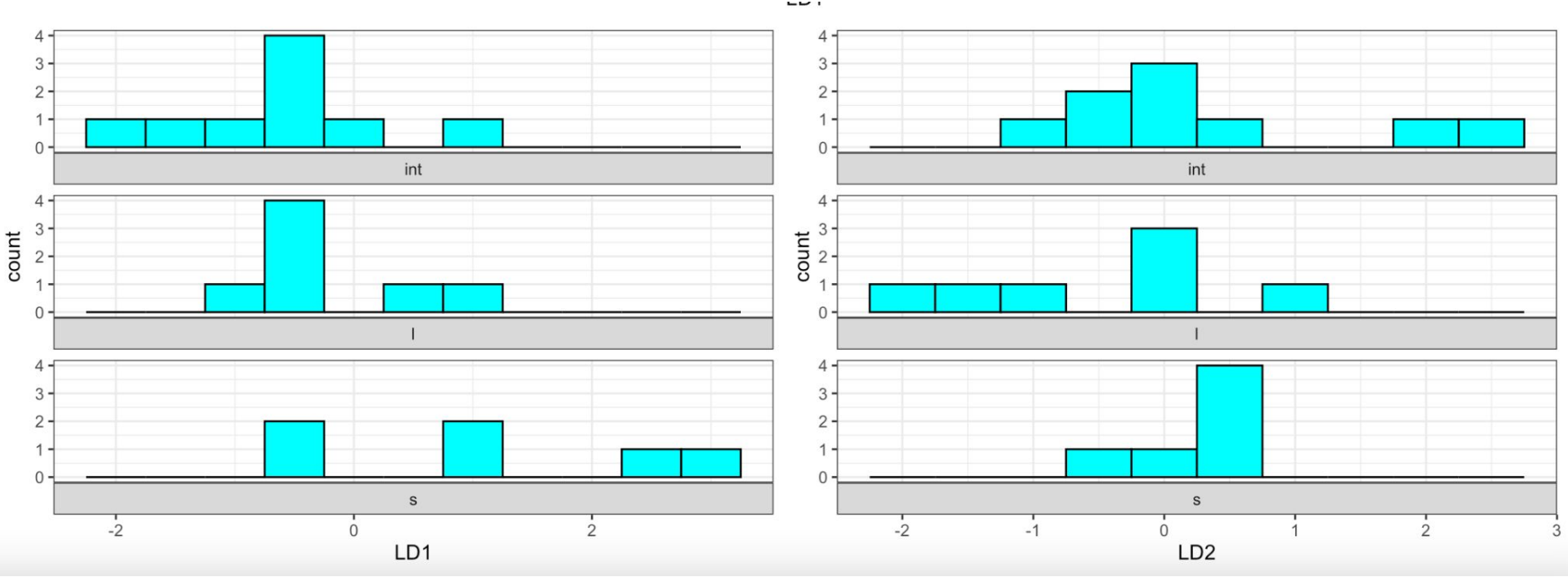
Validating the LDA: does the training dataset predict the grouping in testing dataset?

```
training_lda = lda(data = training_process, group ~ .)
predicted = training_lda %>% predict(testing_process)
print(table(testing_process$group, predicted$class))
```

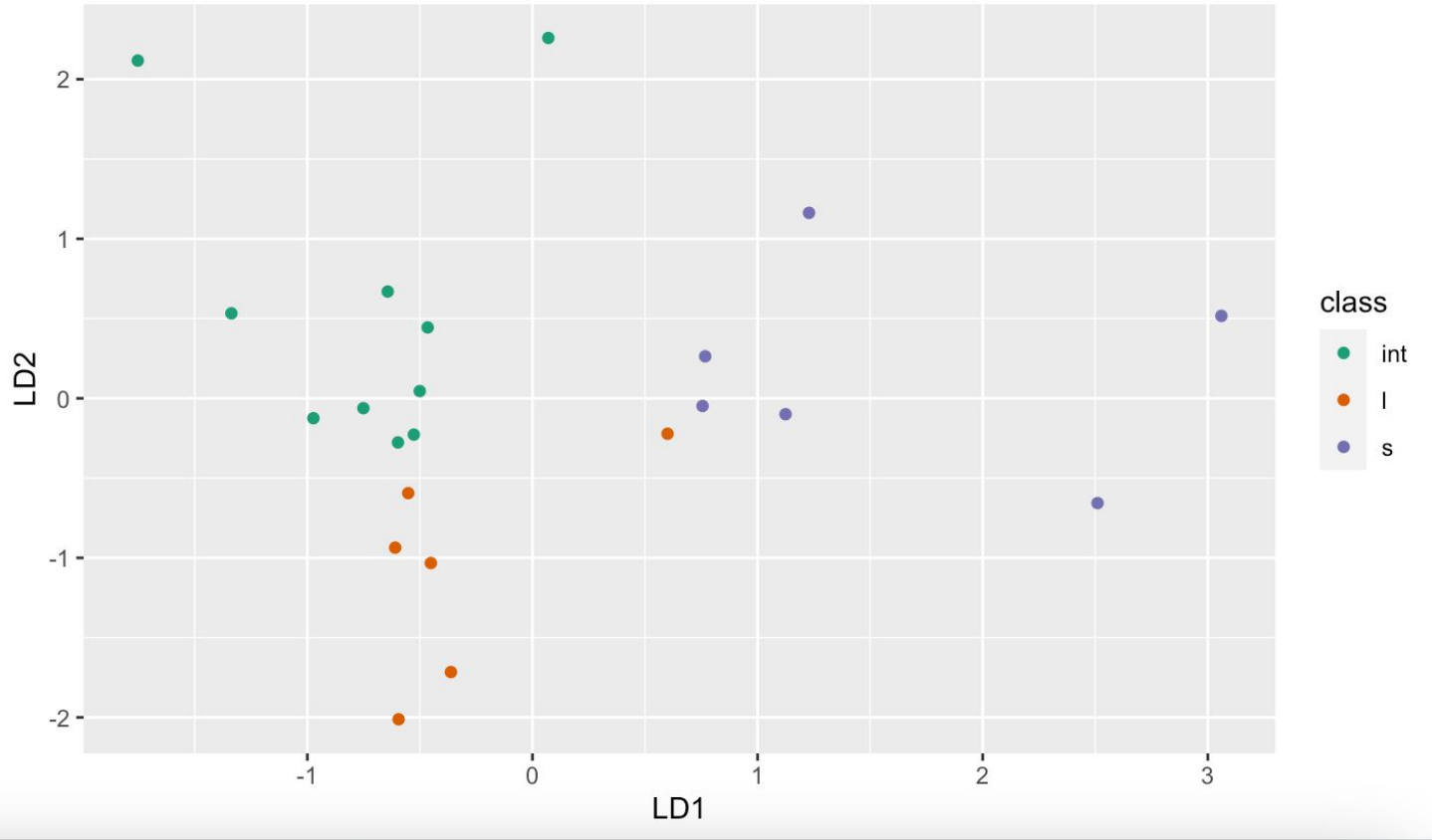
```
     int l s
  int   1 1 2
  l     2 0 1
  s     2 0 1
```

The LDA on training dataset does not predict the groups in the testing dataset very well. The predictive accuracy is 20%.

Plotting the LDs



Plotting the LDA



Which variables contributed most to grouping?

We can gather which variables contributed how much to each discriminant function/LD by looking at the loading/scores of each variable on each LD

lda$scaling displays scores for each variable. By sorting the absolute scores in each LD, you can get a ranking for which variables most explained the grouping

Here, both LD1 and LD2 ranked the variables in the same order. Looks like distance moved and total activity were most responsible for the grouping.

	LD1	LD2
total_act	650.36092	-443.55919
total_dist	-648.38038	442.52466
total_black	509.58009	-353.71603
white_thigmo	231.97374	-160.66499
total_white	231.45582	-160.21054
white_center	142.69416	-98.65417
black_center	92.75756	-64.46677
black_thigmo	85.36472	-59.24759

Discriminant function analyses in other context

LDA assumptions limits its usage. Other DF analyses are available!

Problem 1: You have more potential explanatory variables than you do samples.

Solution: Regularized discriminant analysis (RDA) regularizes data to improve discriminatory power.

Problem 2: You have RNA counts data (non-normal underlying distribution).

Solution: Negative binomial linear discriminant analysis (NBLDA) uses a negative binomial model that more accurately reflects the underlying distribution of counts data.