Discriminant Analysis

Today . . .

- Basic goals of discriminant analysis
- The iris data as an example

The goal of discriminant analysis is to linear combination of variables that can best separate groups.

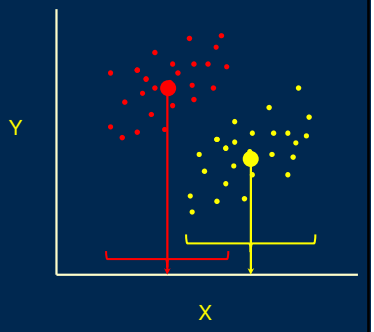
In discriminant analysis, the groups are pre-existing and usually have no particular structure (i.e., their formation is not under experimental control).

Classification is usually an emphasis in discriminant analysis.

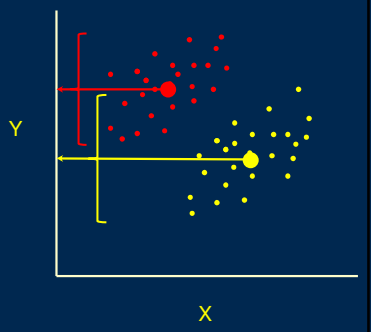
The linear combinations that discriminant analysis constructs will maximize the ratio of between-groups variance to within-groups variance.

If more than one linear combination can be formed, subsequent linear combinations are independent of prior combinations and account for as much remaining group variation as possible (sound familiar?).

These group can be separated by using variable, X, alone. But, the separation is not terrific and many cases would be misclassified if based only on X.

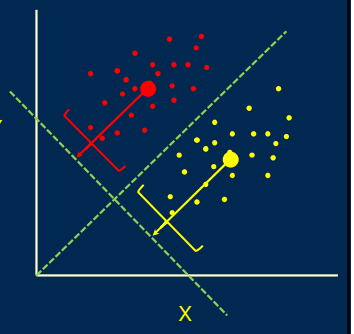


These group can also be separated by using variable, Y, alone. But, here too the separation is not very good and many cases would be misclassified if based only on Y.



Maximum separation uses both variables at the same time (a linear combination).

More variables could be used and they need not be weighted equally.



The Fisher iris data provide familiar territory for introducing basic concepts.

- We know there are three species in the data set and we know that four measures were taken on each flower.
- What is the best linear combination of those four measures for maximizing the separation of the three species?
- How many linear combinations do we need to produce good separation?
- How good is the separation?

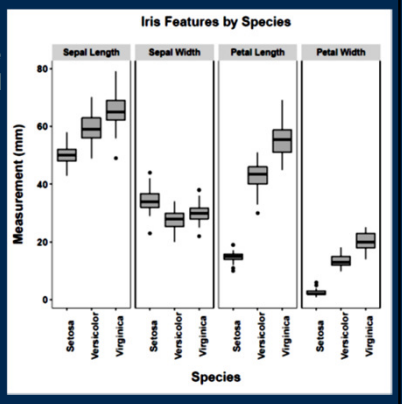
SW = Sepal Width
SL = Sepal Length
PW = Petal Width
PL = Petal Length

$$LC_i = w_1 SW_i + w_2 SL_i + w_3 PW_i + w_4 PL_i$$

Fisher's original formulation identifies weights that maximize:

$$\frac{SS_{Between-Groups}}{SS_{Within-Groups}}$$

Taken individually, the variables (petal measurements in particular) can separate the groups.

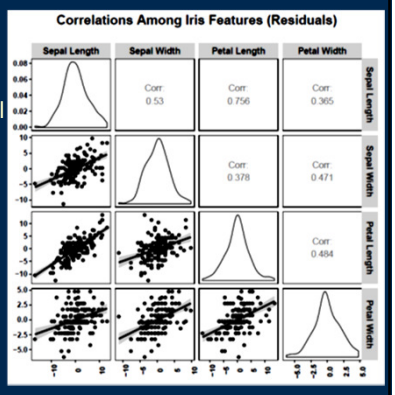



```
Iris_MANOVA <- manova(as.matrix(Iris[, 1:4]) ~ Iris$Species)

summary.aov(Iris_MANOVA)

##  Response Sepal_Length :
##               Df Sum Sq Mean Sq F value Pr(>F)
## Iris$Species   2   6321    3161     119 <2e-16
## Residuals    147   3896      27
##
##  Response Sepal_Width :
##               Df Sum Sq Mean Sq F value Pr(>F)
## Iris$Species   2   1134     567    49.2 <2e-16
## Residuals    147   1696      12
##
##  Response Petal_Length :
##               Df Sum Sq Mean Sq F value Pr(>F)
## Iris$Species   2  43710   21855    1180 <2e-16
## Residuals    147   2722      19
##
##  Response Petal_Width :
##               Df Sum Sq Mean Sq F value Pr(>F)
## Iris$Species   2   8041    4021     960 <2e-16
## Residuals    147    616       4
```

ANOVAs verify the univariate differences for each feature.

The measures are correlated, so their unique contributions will be important to estimate.

Why are these correlations based on residuals (controlling for species)?



Specify the models and save as objects. The objects will have attributes corresponding to most of the important information from the analysis.

More than one package and function will be necessary to get everything we need for interpretation.

```
Iris_LDA <- lda(Species_Num ~ Sepal_Length + Sepal_Width + Petal_Length +
    Petal_Width, data = Iris)
```

```
Iris_MLM <- lm(cbind(Sepal_Length, Sepal_Width, Petal_Length, Petal_Width) ~
    as.factor(Species), data = Iris)
Iris_CDA <- candisc(Iris_MLM, data = Iris)
```

```
Iris_LDA$scaling

##                    LD1      LD2
## Sepal_Length  0.08294 -0.00241
## Sepal_Width   0.15345 -0.21645
## Petal_Length -0.22012  0.09319
## Petal_Width  -0.28105 -0.28392

Iris_CDA$coeffs.raw

##                   Can1     Can2
## Sepal_Length -0.08294 -0.00241
## Sepal_Width  -0.15345 -0.21645
## Petal_Length  0.22012  0.09319
## Petal_Width   0.28105 -0.28392

Iris_CDA$coeffs.std

##                   Can1     Can2
## Sepal_Length -0.4270 -0.01241
## Sepal_Width  -0.5212 -0.73526
## Petal_Length  0.9473  0.40104
## Petal_Width   0.5752 -0.58104
```

As in multiple regression, there are unstandardized (raw) and standardized coefficients (weights). The former are based on centered, but not standardized, variables.

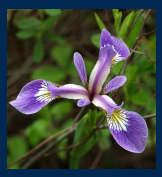
The standardized coefficients show the unique contributions of the variables.

```
Iris_CDA$coeffs.std

##                   Can1     Can2
## Sepal_Length -0.4270 -0.01241
## Sepal_Width  -0.5212 -0.73526
## Petal_Length  0.9473  0.40104
## Petal_Width   0.5752 -0.58104
```

The maximum number of discriminant functions is the smaller of p (number of variables) or g-1 (with g = number of groups). The importance of a particular function will not necessarily be apparent from the weights. Also, the variables are correlated, which can complicate interpretation using the weights alone. The simple correlations of each variable with each function (the structure matrix) should also be examined.

```
Iris_CDA$coeffs.std

##                 Can1     Can2
## Sepal_Length -0.4270 -0.01241
## Sepal_Width  -0.5212 -0.73526
## Petal_Length  0.9473  0.40104
## Petal_Width   0.5752 -0.58104
```

```
Iris_CDA$structure

##                 Can1     Can2
## Sepal_Length  0.7919 -0.21759
## Sepal_Width  -0.5308 -0.75799
## Petal_Length  0.9850 -0.04604
## Petal_Width   0.9728 -0.22290
```



I. Setosa

I. Verginica

I. Versicolor

How many functions to we need in order to separate the species? What is their relative importance? We can use canonical correlations, eigenvalues, and Wilks' Λ to help us make these decisions.

If the discriminating variables represent one set and a group of dummy variables derived from group membership represent another set, then the canonical correlation is the maximum linear relation between weighted combinations from those sets. The canonical correlations can be used to determine the relative discrimination of multiple discriminant functions.

The canonical correlations are related to the eigenvalues (λ_j). These eigenvalues refer to the matrix product of SS_W^{-1} and SS_B.

$$\lambda_j = \frac{r_j^2}{1 - r_j^2}$$

The eigenvalues can be used to calculate the proportion of discrimination that is due to each function:

$$proportion_j = \frac{\lambda_j}{\sum_{d=1}^{D} \lambda_d}$$

The canonical correlations are also related to Wilks' lambda (Λ), used to provide a statistical test for the discriminant functions.

$$\Lambda_j = \prod_{d=j}^{D}(1 - r_d^2)$$

Tests of significance in discriminant analysis are made in a step-wise fashion. First, the entire set of functions is tested for significance. If this test is significant, then we conclude that there is significant discrimination possible.

Then the first (and most important) function is excluded and the remainder are tested. If this is not significant, then the first function was the only source of discrimination.

If the remainder is significant, then the first and at least the second are significant sources of discrimination.

```
Canonical Discriminant Analysis for as.factor(Species):

  CanRsq Eigenvalue Difference Percent Cumulative
1  0.970     32.192       31.9  99.121       99.1
2  0.222      0.285       31.9   0.879      100.0

Test of H0: The canonical correlations in the
current row and all that follow are zero

  LR test stat approx F numDF denDF     Pr(> F)
1        0.023    199.1     8   288     < 2e-16
2        0.778     13.8     3   145 0.000000058
```

Both discriminant functions are significant but the first is far more important.

Λ_1 = .023 = (1-.970)(1-.222) Λ_2 = (1-.222)

We assume in discriminant analysis that the separate group variance-covariance matrices are homogeneous. This assumption underlies tests used to determine the number of significant functions. Box's test can be used to test this assumption.

```
 Box's M-test for Homogeneity of Covariance Matrices

data:  Iris[, 1:4]
Chi-Sq (approx.) = 140, df = 20, p-value <2e-16
```

The test implies that the pooled variance-covariance matrix does not represent the groups well:

	Sepal_Length	Sepal_Width	Petal_Length	Petal_Width
Sepal_Length	26.501	9.272	16.751	3.840
Sepal_Width	9.272	11.539	5.524	3.271
Petal_Length	16.751	5.524	18.519	4.267
Petal_Width	3.840	3.271	4.267	4.188

$Setosa

	Sepal_Length	Sepal_Width	Petal_Length	Petal_Width
Sepal_Length	12.425	9.9216	1.6355	1.0331
Sepal_Width	9.922	14.3690	1.1698	0.9298
Petal_Length	1.636	1.1698	3.0159	0.6069
Petal_Width	1.033	0.9298	0.6069	1.1106

$Versicolor

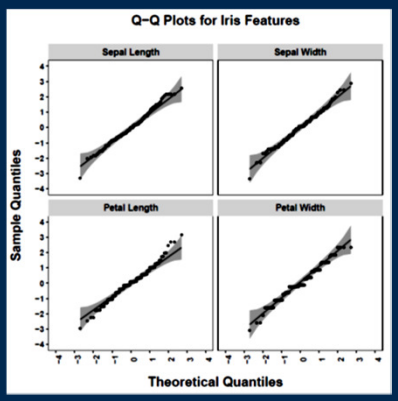
	Sepal_Length	Sepal_Width	Petal_Length	Petal_Width
Sepal_Length	26.643	8.518	18.290	5.578
Sepal_Width	8.518	9.847	8.265	4.120
Petal_Length	18.290	8.265	22.082	7.310
Petal_Width	5.578	4.120	7.310	3.911

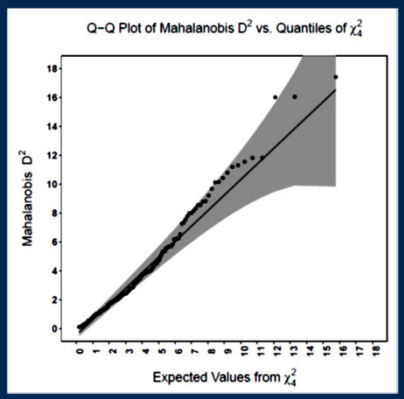
$Virginica

	Sepal_Length	Sepal_Width	Petal_Length	Petal_Width
Sepal_Length	40.434	9.376	30.329	4.909
Sepal_Width	9.376	10.400	7.138	4.763
Petal_Length	30.329	7.138	30.459	4.882
Petal_Width	4.909	4.763	4.882	7.543

Box's test indicates that these are not equal. An alternative form of the analysis (quadratic) can be used that does not assume homogeneity.

We also assume multivariate normality for the significance tests as well as the classification part of discriminant analysis. The tests are performed on the residualized data so that species differences do not affect the results. Note that a violation of multivariate normality will also affect the test of homogeneity of covariance matrices.






```
mvn(Iris[, 11:14], mvnTest = "mardia")

## $multivariateNormality
##              Test         Statistic              p value Result
## 1 Mardia Skewness 31.8480630655843  0.0449444331703922     NO
## 2 Mardia Kurtosis 3.28196485281224 0.00103086453506429     NO
## 3             MVN              <NA>                  <NA>     NO
##
## $univariateNormality
##           Test  Variable Statistic   p value Normality
## 1 Shapiro-Wilk   SL_R       0.9879    0.2189    YES
## 2 Shapiro-Wilk   SW_R       0.9895    0.3230    YES
## 3 Shapiro-Wilk   PL_R       0.9811    0.0368    NO
## 4 Shapiro-Wilk   PW_R       0.9722    0.0039    NO
```

The evidence for non-normality is not overwhelming.

The quality of the discriminant function can be assessed by how well the cases can be classified into their known groups, and, how well new cases can be classified. Three increasingly rigorous classifications can be examined:

1. All cases
2. Leave-One-Out (Jackknife)
3. Separate sample cross-validation.

Classification can be done in several ways. The most common approach uses a Bayesian model that takes prior probabilities into account. Cases are classified into the group for which they have the highest posterior probability.

```
Iris_Predicted <- predict(Iris_LDA)
Iris_Predicted$class

##   [1] 1 3 2 3 2 1 3 2 3 1 2 3 3 2 3 3 3 1 2 3 3 2 3 3 3 1 3 2 2 2
##  [31] 1 3 2 3 3 1 1 2 3 1 3 1 2 1 3 3 1 2 2 3 1 1 3 1 1 1 3 3 1 1
##  [61] 1 2 3 1 1 2 2 1 1 2 2 1 1 3 3 3 2 3 1 1 3 3 3 3 2 2 2 1 1 3
##  [91] 3 1 2 2 2 1 1 2 2 2 1 1 3 2 3 2 1 1 3 2 3 3 1 2 2 1 2 2 2 2
## [121] 2 2 3 3 1 1 3 3 2 2 2 3 3 2 1 1 1 3 1 1 2 2 2 1 1 1 3 2 3 1
## Levels: 1 2 3
```

```
Iris_Predicted$posterior

##            1         2         3
## 1   1.000e+00 2.322e-20 4.242e-40
## 2   1.320e-45 3.014e-06 1.000e+00
## 3   4.214e-23 9.956e-01 4.410e-03
## 4   6.571e-45 1.181e-06 1.000e+00
## 5   1.284e-28 7.294e-01 2.706e-01
```

Classification by posterior probabilities is outstanding for the iris data.

```
Iris_Fisher$confusion

##             predicted
## original     Setosa Versicolor Virginica
##   Setosa         50          0         0
##   Versicolor      0         48         2
##   Virginica       0          1        49
```

The jackknife procedure will leave each case out in turn, estimate the discriminant analysis with the remaining cases, and then use that information to classify the left-out case. This approach insures that each case is classified with information it did not contribute to in the estimation.

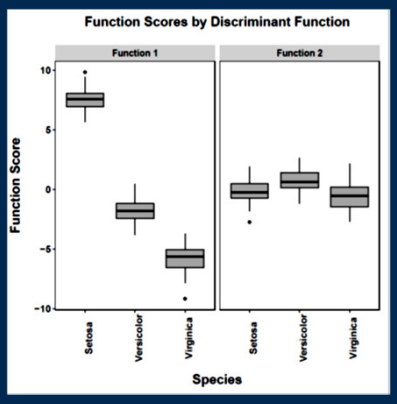
```
Iris_Jack <- lda(Species ~ Sepal_Length + Sepal_Width + Petal_Length +
    Petal_Width, data = Iris, CV = TRUE)
table(Original = Iris$Species_Num, Predicted = Iris_Jack$class)

##          Predicted
## Original Setosa Versicolor Virginica
##        1     50          0         0
##        2      0         48         2
##        3      0          1        49
```

The most convincing cross-validation uses part of the sample (or a separate sample) to estimate the discriminant functions and then applies that solution to the remaining cases. Here the sample is split into random halves. The first sample (called the training set) is used to derive the discriminant functions. Those functions are then used on the second set to classify cases.

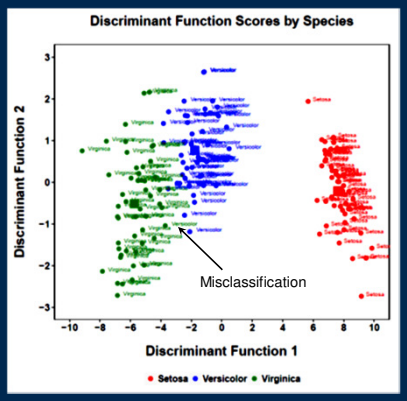
```
training_sample <- sample(1:150, 75)

Iris_Train <- lda(Species ~ Sepal_Length + Sepal_Width + Petal_Length +
    Petal_Width, data = Iris, CV = FALSE, subset = training_sample)

Iris_Predict <- predict(Iris_Train, newdata = Iris[-training_sample,
    ])
Iris_Original <- as.data.frame(Iris[-training_sample, 5])
Iris_Cross <- cbind(Iris_Original, Iris_Predict$class)
names(Iris_Cross) <- c("Original_Species", "Predicted_Species")
table(Original = Iris_Cross$Original_Species, Predicted = Iris_Cross$Predicted_Species)

##             Predicted
## Original     Setosa Versicolor Virginica
##   Setosa         25          0         0
##   Versicolor      0         24         0
##   Virginica       0          1        25
```

The discriminant functions from the first sample correctly classify 98.7% of the cases in the second sample.

The results of a discriminant analysis can be visualized in a number of ways. When comparing the results across discriminant functions, using a common scale can show the different discrimination power.



The superior separation with the first discriminant function is evident.



Color is the predicted group. Label is the actual group.

Note that different scaling is used here to reduce clutter.

Next time . . .

- A closer look at classification
- Violation of assumptions