A researcher collected data on two external features for two (known) sub-species of an insect. She can use discriminant analysis to find linear combinations of the features that best distinguish the sub-species. The analysis can then be used to classify insects with unknown sub-species origin into one of the two sub-species based on their external features.

To see how this might be done, consider the following data plot. Can1 is the linear combination of the two features that best distinguishes or discriminates the two sub-species. The value of Can1 could be used to classify insects into one of the two groups, as illustrated.
The method generalizes to more than two features and sub-species.
16.1 Canonical Discriminant Analysis

While there’s a connection between canonical discriminant analysis and canonical correlation, I prefer to emphasize the connection between canonical discriminant analysis and MANOVA because these techniques are essentially identical.

Assume that you have representative samples from $k$ groups, strata, or subpopulations. Each selected individual is measured on $p$ features (measurements) $X_1, X_2, \ldots, X_p$. As in MANOVA, canonical discriminant analysis assumes you have independent samples from multivariate normal populations with identical variance-covariance matrices.

Canonical discriminant analysis computes $r = \min(p, k - 1)$ linear combinations of the features with the following properties. The first linear combination, called the first linear discriminant function

$$\text{Can}_1 = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p$$

gives the most significant $F$-test for a null hypothesis of no group differences in a one-way ANOVA, among all linear combinations of the features. The second linear combination or the second linear discriminant function:

$$\text{Can}_2 = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p$$

gives the most significant $F$-test for no group differences in a one-way ANOVA, among all linear combinations of the features that are uncorrelated (adjusting for groups) with $\text{Can}_1$. In general, the $j^{th}$ linear combination $\text{Can}_j$ ($j = 1, 2, \ldots, r$) gives the most significant $F$-test for no group differences in a one-way ANOVA, among all linear combinations of the features that are uncorrelated with $\text{Can}_1, \text{Can}_2, \ldots, \text{Can}(j - 1)$.

The coefficients in the canonical discriminant functions can be multiplied by a constant, or all the signs can be changed (that is, multiplied by the constant $-1$), without changing their properties or interpretations.
16.2 Example: Owners of riding mowers

The manufacturer of a riding lawn mower wishes to identify the best prospects for buying their product using data on the incomes ($X_1$) and lot sizes ($X_2$) of homeowners (Johnson and Wichern, 1988). The data below are the incomes and lot sizes from independent random samples of 12 current owners and 12 non-owners of the mowers.

```r
library(tidyverse)

# load ada functions
source("ada_functions.R")

### Example: Riding mowers
dat_mower <-
  read_table2(
    "http://statacumen.com/teach/ADA2/notes/ADA2_notes_Ch16_mower.dat"
  )

## Parsed with column specification:
## cols(
##   income = col_double(),
##   lotsize = col_double(),
##   owner = col_character()
## )

# income = income in £1000
# lotsize = lot size in 1000 sq ft
# owner = nonowners or owners
str(dat_mower)
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 24 obs. of 3 variables:
## $ income : num 20 28.5 21.6 20.5 29 36.7 36 27.6 23 31 ... 
## $ lotsize: num 9.2 8.4 10.8 10.4 11.8 9.6 8.8 11.2 10 10.4 ... 
## $ owner  : chr "owner" "owner" "owner" "owner" "owner" "owner" "owner" ... 
## - attr(*, "spec")=
##   .. cols(
##   ..   income = col_double(),
##   ..   lotsize = col_double(),
##   ..   owner = col_character()
##   .. )
```
### 16.2: Example: Owners of riding mowers

<table>
<thead>
<tr>
<th>income</th>
<th>lotsize</th>
<th>owner</th>
</tr>
</thead>
<tbody>
<tr>
<td>20.00</td>
<td>9.20</td>
<td>owner</td>
</tr>
<tr>
<td>28.50</td>
<td>8.40</td>
<td>owner</td>
</tr>
<tr>
<td>21.60</td>
<td>10.80</td>
<td>owner</td>
</tr>
<tr>
<td>20.50</td>
<td>10.40</td>
<td>owner</td>
</tr>
<tr>
<td>29.00</td>
<td>11.80</td>
<td>owner</td>
</tr>
<tr>
<td>36.70</td>
<td>9.60</td>
<td>owner</td>
</tr>
<tr>
<td>36.00</td>
<td>8.80</td>
<td>owner</td>
</tr>
<tr>
<td>27.60</td>
<td>11.20</td>
<td>owner</td>
</tr>
<tr>
<td>23.00</td>
<td>10.00</td>
<td>owner</td>
</tr>
<tr>
<td>31.00</td>
<td>10.40</td>
<td>owner</td>
</tr>
<tr>
<td>17.00</td>
<td>11.00</td>
<td>owner</td>
</tr>
<tr>
<td>27.00</td>
<td>10.00</td>
<td>owner</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>income</th>
<th>lotsize</th>
<th>owner</th>
</tr>
</thead>
<tbody>
<tr>
<td>25.00</td>
<td>9.80</td>
<td>nonowner</td>
</tr>
<tr>
<td>17.60</td>
<td>10.40</td>
<td>nonowner</td>
</tr>
<tr>
<td>21.60</td>
<td>8.60</td>
<td>nonowner</td>
</tr>
<tr>
<td>14.40</td>
<td>10.20</td>
<td>nonowner</td>
</tr>
<tr>
<td>28.00</td>
<td>8.80</td>
<td>nonowner</td>
</tr>
<tr>
<td>19.80</td>
<td>8.00</td>
<td>nonowner</td>
</tr>
<tr>
<td>22.00</td>
<td>9.20</td>
<td>nonowner</td>
</tr>
<tr>
<td>15.80</td>
<td>8.20</td>
<td>nonowner</td>
</tr>
<tr>
<td>11.00</td>
<td>9.40</td>
<td>nonowner</td>
</tr>
<tr>
<td>17.00</td>
<td>7.00</td>
<td>nonowner</td>
</tr>
<tr>
<td>16.40</td>
<td>8.80</td>
<td>nonowner</td>
</tr>
<tr>
<td>21.00</td>
<td>7.40</td>
<td>nonowner</td>
</tr>
</tbody>
</table>

```r
library(ggplot2)
p <- ggplot(dat_mower, aes(x = income, y = lotsize, shape = owner, colour = owner))
p <- p + geom_point(size = 3)
p <- p + scale_y_continuous(limits = c(0, 15))
p <- p + scale_x_continuous(limits = c(0, 40))
p <- p + coord_fixed(ratio = 1) # square axes (for perp lines)
p <- p + xlab("Income in $1000")
p <- p + ylab("Lot size in 1000 sq ft")
print(p)
```

```r
library(GGally)
p <-
  ggpairs(
    rev(dat_mower)
    , mapping = ggplot2::aes(colour = owner, alpha = 0.5)
    , progress = FALSE
  )
print(p)
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```
Although the two groups overlap, the owners tend to have higher incomes and larger lots than the non-owners. Income seems to distinguish owners and non-owners better than lot size, but both variables seem to be useful for discriminating between groups.

Qualitatively, one might classify prospects based on their location relative to a roughly vertical line on the scatter plot. A discriminant analysis gives similar results to this heuristic approach because the Can1 scores will roughly correspond to the projection of the two features onto a line perpendicular to the hypothetical vertical line. `candisc()` computes one discriminant function here because \( p = 2 \) and \( k = 2 \) gives \( r = \min(p, k - 1) = \min(2, 1) = 1 \).

Below we first fit a `lm()` and use that object to compare populations. First we compare using univariate ANOVAs. The p-values are for one-way ANOVA comparing owners to non-owners and both income and lotsize features are im-
important individually for distinguishing between the groups.

```r
# first fit lm() with formula = continuous variables ~ factor variables
lm_mower <-
  lm(
    cbind(income, lotsize) ~ owner,
    data = dat_mower
  )

# univariate ANOVA tests
summary(lm_mower)
```

## Response income :

### Call:
`lm(formula = income ~ owner, data = dat_mower)`

### Residuals:
```
Min 1Q Median 3Q Max
-9.4917 -3.8021 0.5875 2.5979 10.2083
```

### Coefficients:
```
(Intercept) 19.133 1.601 11.954 4.28e-11 ***
ownerowner 7.358 2.264 3.251 0.00367 **
```

### Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

### Residual standard error: 5.545 on 22 degrees of freedom

### Multiple R-squared: 0.3245, Adjusted R-squared: 0.2938

### F-statistic: 10.57 on 1 and 22 DF, p-value: 0.003665

## Response lotsize :

### Call:
`lm(formula = lotsize ~ owner, data = dat_mower)`

### Residuals:
```
Min 1Q Median 3Q Max
-1.81667 -0.66667 -0.01667 0.71667 1.66667
```

### Coefficients:
```
(Intercept) 8.8167 0.2984 29.55 < 2e-16 ***
ownerowner 1.3167 0.4220 3.12 0.00498 **
```

### Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

### Residual standard error: 1.034 on 22 degrees of freedom

UNM, Stat 428/528 ADA2
Second, the MANOVA indicates the multivariate means are different indicating both income and lotsize features taken together are important for distinguishing between the groups.

```r
# test whether the multivariate means of the two populations are different
library(car)
man_mower <- Manova(lm_mower)
summary(man_mower)
```

Finally, we fit the canonical discriminant function with `candisc()`. The LR (likelihood ratio) p-values below correspond to tests of no differences between groups on the canonical discriminant functions. There is only one canonical discriminant function here. The tests of no differences based on the first canonical discriminant function is equivalent to Roy's MANOVA test.

```r
# perform canonical discriminant analysis
library(candisc)
can_mower <- candisc(lm_mower)
```
can_mower

## Canonical Discriminant Analysis for owner:
##
## CanRsq Eigenvalue Difference Percent Cumulative
## 1  0.5386  1.1673  100  100
##
## 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  0.4614  2

The objects available from the \texttt{candisc()} object are named below, and we’ll soon use a few. There are also a few plots available, but I’ll be creating other plots shortly.

\texttt{names(can_mower)} # list of objects in \texttt{can_mower}

## [1] ”dfh” ”dfe” ”eigenvalues” ”canrsq”
## [5] ”pct” ”rank” ”ndim” ”means”
## [9] ”factors” ”term” ”terms” ”coeffs.raw”
## [13] ”coeffs.std” ”structure” ”scores”

\texttt{plot(can_mower)} # boxplots with proportional contribution of each variable to Can1

### can also plot 2D plots when have more than two groups (will use later)

\texttt{library(heplots)}
\texttt{heplot(can_mower, scale=6, fill=TRUE)}
\texttt{heplot3d(can_mower, scale=6, fill=TRUE)}

The raw canonical coefficients define the canonical discriminant variables and are identical to the feature loadings in a one-way MANOVA, except for an unimportant multiplicative factor. Only Can1 is generated here.

UNM, Stat 428/528 ADA2
The means output gives the mean score on the canonical discriminant variables by group, after centering the scores to have mean zero over all groups. These are in order of the owner factor levels (nonowner, owner).

The linear combination of income and lotsize that best distinguishes owners from non-owners

$$\text{Can1} = -0.1453 \ \text{INCOME} + -0.759 \ \text{LOTSIZE}$$

is a weighted average of income and lotsize.

In the scatterplot below, Can1 is the direction indicated by the dashed line.
The standardized coefficients (use the pooled within-class coefficients) in-
dicate the relative contributions of the features to the discrimination. The standardized coefficients are roughly equal, which suggests that income and lotsize contribute similarly to distinguishing the owners from non-owners.

```
can_mower$coeffs.std
## Can1
## income -0.8058419
## lotsize -0.7845512
```

The p-value of 0.0004 on the likelihood ratio test indicates that Can1 strongly distinguishes between owners and non-owners. This is consistent with the separation between owners and non-owners in the boxplot of Can1 scores.

I noted above that Can1 is essentially the same linear combination given in a MANOVA comparison of owners to non-owners. Here is some `Manova()` output to support this claim. The MANOVA test p-values agree with the `candisc` output (as we saw earlier). The first characteristic vector from the MANOVA is given here.

```
## For Roy's characteristic Root and vector
H <- man_mower$SSP$owner  # H = hypothesis matrix
E <- man_mower$SSPE       # E = error matrix
# characteristic roots of (E inverse * H)
EinvH <- solve(E) %*% H   # solve() computes the matrix inverse
ev <- eigen(EinvH)       # eigenvalue/eigenvectors

ev
## $values
## [1] 1.167337 0.000000
##
## $vectors
## [,1] [,2]
## [1,] 0.1880613 -0.1761379
## [2,] 0.9821573 0.9843655

mult_char_can_disc <- can_mower$coeffs.raw[1] / ev$vectors[1,1]
mult_char_can_disc
## [1] -0.7728352
```

The first canonical discriminant function is obtained by multiplying the first characteristic vector given in MANOVA by -0.7728:

```
Can1 = -0.1453 INCOME + -0.759 LOTSIZE
      = -0.7728 (0.1881 INCOME + 0.9822 LOTSIZE)
```
Fisher’s iris data consists of samples of 50 flowers from each of three species of iris: Setosa, Versicolor, and Virginica. Four measurements (in mm) were taken on each flower: sepal length, sepal width, petal length, and petal width.

The plots show big differences between Setosa and the other two species. The differences between Versicolor and Virginica are smaller, and appear to be mostly due to differences in the petal widths and lengths.

```r
### Example: Fisher’s iris data
# The "iris" dataset is included with R in the library(datasets)
data(iris)
str(iris)

# 'data.frame': 150 obs. of 5 variables:
# $ Sepal.Length: num 5.1 4.9 4.7 4.6 5.4 4.6 5 4.4 4.9 ... 
# $ Sepal.Width : num 3.5 3.2 3.1 3.6 3.9 3.4 3.4 3.4 2.9 3.1 ... 
# $ Petal.Length: num 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.5 1.5 ... 
# $ Petal.Width : num 0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ... 
# $ Species : Factor w/ 3 levels "setosa","versicolor",...: 1 1 1 1 1 1 1 1 1 1 ... 

# Scatterplot matrix
library(ggplot2)
library(GGally)
p <-
  ggpairs(
    iris[,c(5,1,2,3,4)]
    , mapping = ggplot2::aes(colour = Species, alpha = 0.5)
    , progress = FALSE
  )
print(p)
```

UNM, Stat 428/528 ADA2
## parallel coordinate plot

```r
library(ggplot2)
library(GGally)

# univariate min/max scaling
p1 <- ggparcoord(
  data = iris,
  columns = 1:4,
  groupColumn = 5,
  order = "anyClass",
  scale = "uniminmax"  # "uniminmax", "globalminmax"
  # showPoints = FALSE,
  title = "uniminmax scaling",
  alphaLines = 1/3
  #, shadeBox = "white"
)
```

Cor : −0.118
setosa: 0.743
versicolor: 0.526
virginica: 0.457
Cor : 0.872
setosa: 0.267
versicolor: 0.754
virginica: 0.864
Cor : −0.428
setosa: 0.178
versicolor: 0.561
virginica: 0.401
Cor : 0.818
setosa: 0.278
versicolor: 0.546
virginica: 0.281
Cor : −0.366
setosa: 0.233
versicolor: 0.664
virginica: 0.538
Cor : 0.963
setosa: 0.332
versicolor: 0.787
virginica: 0.322
\texttt{candisc} was used to discriminate among species. There are $k = 3$ species and $p = 4$ features, so the number of discriminant functions is 2 (the minimum of 4 and $3 - 1$).
# first fit lm() with formula = continuous variables ~ factor variables
lm_iris <-
  lm(
    cbind(Sepal.Length, Sepal.Width, Petal.Length, Petal.Width) ~ Species
    , data = iris
  )

## univariate ANOVA tests
#summary(lm_iris)
## test whether the multivariate means of the two populations are different
#library(car)
#man_mower <- Manova(lm_iris)
#summary(man_mower)
# perform canonical discriminant analysis

library(candisc)
can_iris <- candisc(lm_iris)
can_iris$coeffs.raw

## Can1 Can2
## Sepal.Length -0.8293776 0.02410215
## Sepal.Width -1.5344731 2.16452123
## Petal.Length 2.2012117 -0.93192121
## Petal.Width 2.8104603 2.83918785

Can1 is a comparison of petal and sepal measurements (from Raw Canonical Coefficients):

Can1 = −0.8294 sepalL + −1.534 sepalW + 2.201 petalL + 2.81 petalW.

Can2 is not easily interpreted, though perhaps a comparison of lengths and widths ignoring sepalL:

Can2 = 0.0241 sepalL + 2.165 sepalW + −0.9319 petalL + 2.839 petalW.

The canonical directions provide a maximal separation the species. Two lines across Can1 will provide a classification rule.
There are significant differences among species on both discriminant functions; see the p-values under the likelihood ratio tests. Of course, Can1 produces the largest differences — the overlap among species on Can1 is small. Setosa has the lowest Can1 scores because this species has the smallest petal measurements relative to its sepal measurements. Virginica has the highest Can1 scores.

There are significant differences among species on both discriminant functions; see the p-values under the likelihood ratio tests. Of course, Can1 produces the largest differences — the overlap among species on Can1 is small. Setosa has the lowest Can1 scores because this species has the smallest petal measurements relative to its sepal measurements. Virginica has the highest Can1 scores.

can_iris
##
## Canonical Discriminant Analysis for Species:

UNM, Stat 428/528 ADA2
## Ch 16: Discriminant Analysis

### CanRsq Eigenvalue Difference Percent Cumulative

<table>
<thead>
<tr>
<th></th>
<th>CanRsq</th>
<th>Eigenvalue</th>
<th>Difference</th>
<th>Percent</th>
<th>Cumulative</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.96987</td>
<td>32.19193</td>
<td>31.907</td>
<td>99.12126</td>
<td>99.121</td>
</tr>
<tr>
<td>2</td>
<td>0.22203</td>
<td>0.28539</td>
<td>31.907</td>
<td>0.87874</td>
<td>100.000</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>LR test stat</th>
<th>approx F</th>
<th>numDF</th>
<th>denDF</th>
<th>Pr(&gt; F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.02344</td>
<td>199.145</td>
<td>8</td>
<td>&lt; 2.2e-16 ***</td>
</tr>
<tr>
<td>2</td>
<td>0.77797</td>
<td>13.794</td>
<td>3</td>
<td>5.794e-08 ***</td>
</tr>
</tbody>
</table>

---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

### Questions:

1. What is the most striking feature of the plot of the Can1 scores?
2. Does the assumption of equal population covariance matrices across species seem plausible?
3. How about multivariate normality?

### Covariance matrices by species

```r
by(iris[,1:4], iris$Species, cov)
```

#### iris$Species: setosa
```r
# iris$Species: setosa
## Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length 0.12424898 0.099216327 0.016355102 0.010330612
## Sepal.Width 0.09921633 0.143689796 0.011697959 0.009297959
## Petal.Length 0.01635510 0.011697959 0.030159184 0.006069388
## Petal.Width 0.01033061 0.009297959 0.006069388 0.011106122
```

#### iris$Species: versicolor
```r
# iris$Species: versicolor
## Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length 0.26643265 0.08518367 0.18289796 0.05577959
## Sepal.Width 0.08518367 0.09846939 0.08265306 0.04120408
## Petal.Length 0.18289796 0.08265306 0.22081633 0.07310204
## Petal.Width 0.05577959 0.04120408 0.07310204 0.03910612
```

#### iris$Species: virginica
```r
# iris$Species: virginica
## Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length 0.40434286 0.09376327 0.30328980 0.04909388
## Sepal.Width 0.09376327 0.10400408 0.07137959 0.04762857
## Petal.Length 0.30328980 0.07137959 0.30458776 0.04882449
## Petal.Width 0.04909388 0.04762857 0.04882449 0.07543265
```

# Test multivariate normality using the Shapiro-Wilk test for multivariate normality

```r
library(mvnormtest)
```

---

Prof. Erik B. Erhardt
# The data needs to be transposed t() so each variable is a row
# with observations as columns.

mshapiro.test(t(iris[iris$Species == "setosa" , 1:4]))
##
## Shapiro-Wilk normality test
##
## data: Z
## W = 0.95878, p-value = 0.07906

mshapiro.test(t(iris[iris$Species == "versicolor", 1:4]))
##
## Shapiro-Wilk normality test
##
## data: Z
## W = 0.93043, p-value = 0.005739

mshapiro.test(t(iris[iris$Species == "virginica" , 1:4]))
##
## Shapiro-Wilk normality test
##
## data: Z
## W = 0.93414, p-value = 0.007955

# Graphical Assessment of Multivariate Normality
par(mfrow=c(1,3))
f_mnv_norm_qqplot(iris[iris$Species == "setosa" , 1:4], "setosa")
f_mnv_norm_qqplot(iris[iris$Species == "versicolor", 1:4], "versicolor")
f_mnv_norm_qqplot(iris[iris$Species == "virginica" , 1:4], "virginica")
par(mfrow=c(1,1))