

STA30005:

MULTIVARIATE ANALYSIS

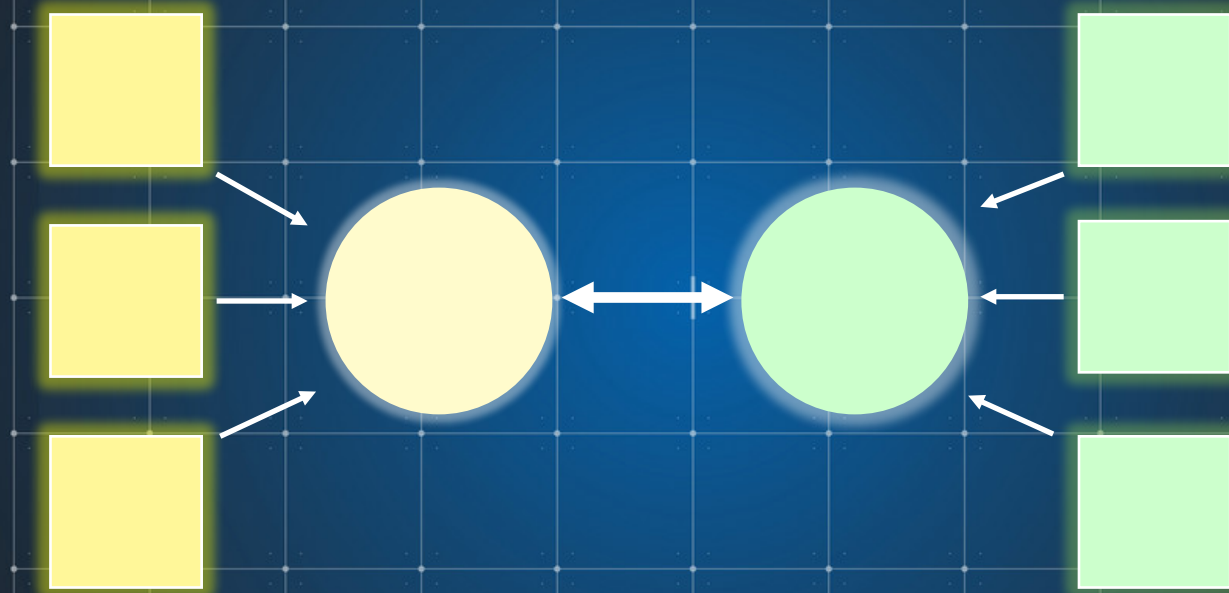


WEEK 10:

CANONICAL CORRELATION

Tutorial Slides

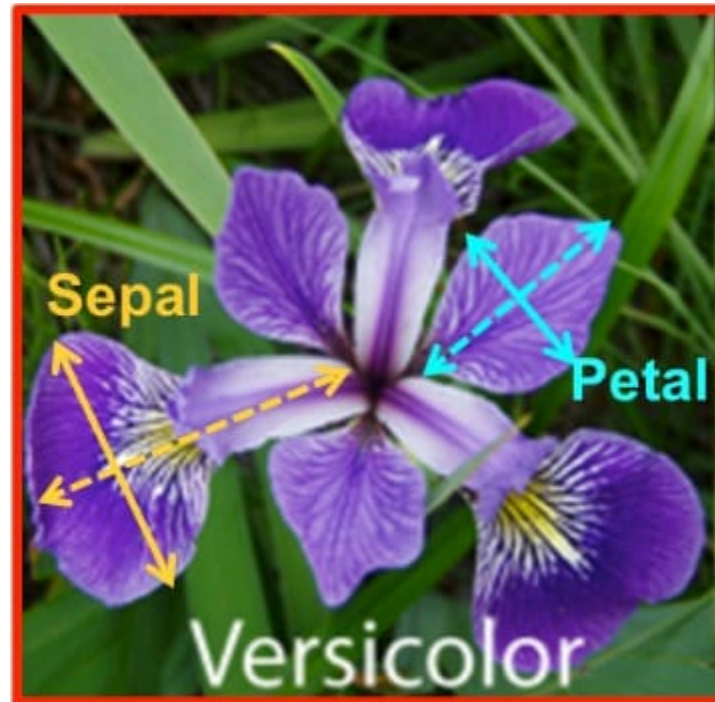
STA30005: Multivariate Analysis



THE DATA

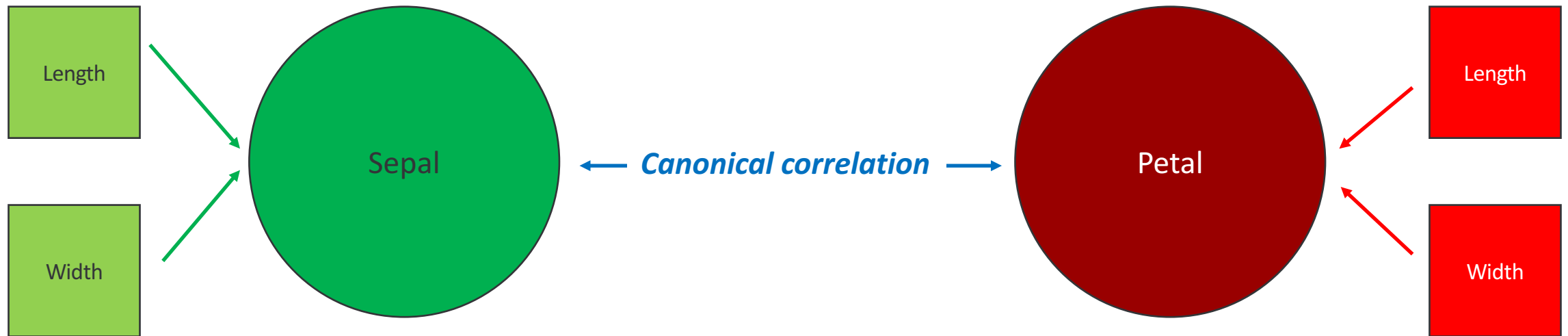
The data file [iris](#) contains the measurements for 150 iris flowers regarding the length and width of their sepals and petals.

Naturally, the sepal length and width should be related (as well as the petal length and width). In other words, there is collinearity among the variables. This makes multiple regression inappropriate and other methods (i.e. canonical correlations) would be required.



THE MODEL

In this model we believe there are two variates (Sepal and Petal), both of which have two variables: length and width. Conduct a canonical correlation on the iris data and determine the main elements for this model



CANONICAL CORRELATION

Data / Package Preparation

Bivariate Correlations

Canonical Correlations

Significance Testing

Cross-loadings and vis.

Load the data (this is a base data set for R)

```
data("iris")
```

Show the data

```
View(iris)
```

Define the first variate (sepal) based on columns 1 and 2

```
sepal<-iris[,1:2]
```

Define the second variate (petal) based on columns 3 and 4

```
petal<-iris[,3:4]
```

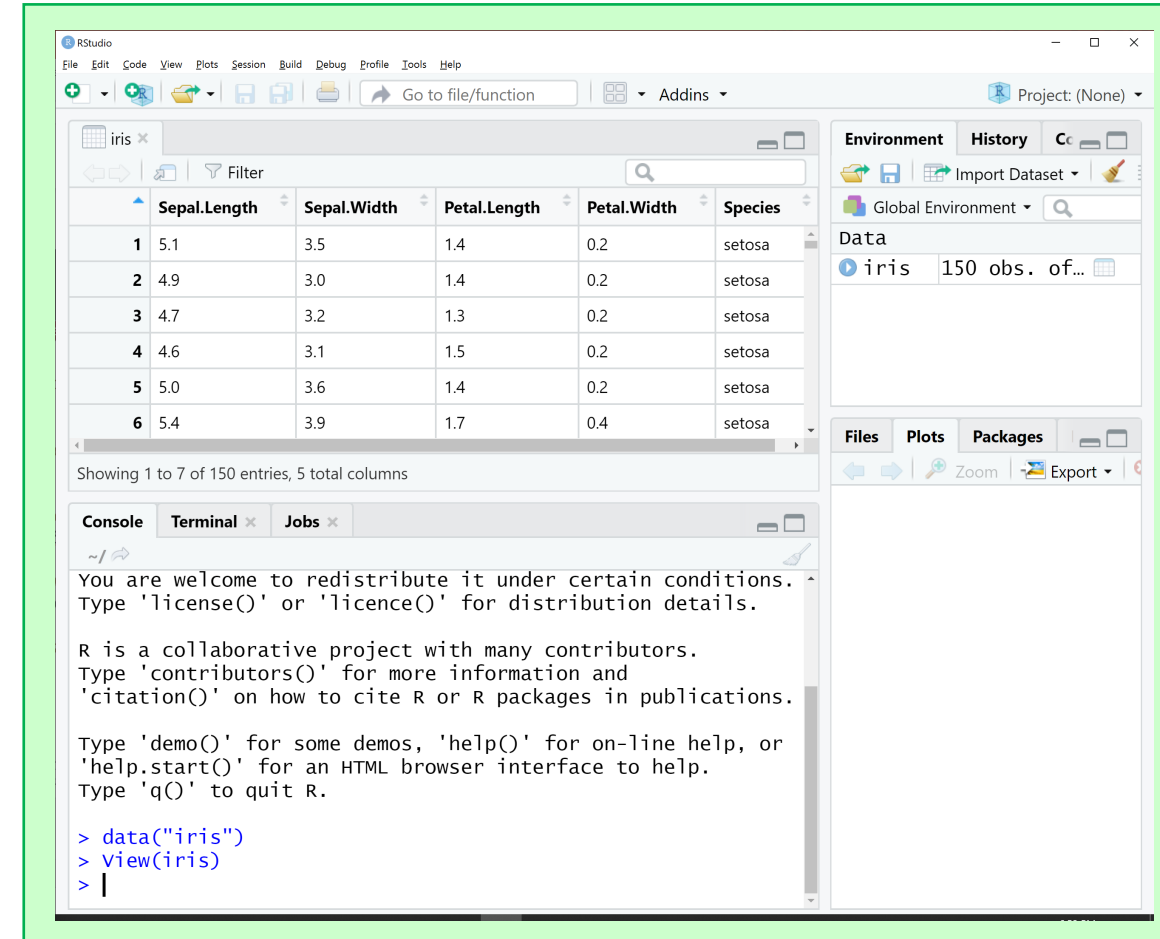
Install / Load the required packages

```
install.packages("yacca")
```

```
library(yacca)
```

```
install.packages("CCA")
```

```
library(CCA)
```



CANONICAL CORRELATION

Data / Package Preparation

Bivariate Correlations

Canonical Correlations

Significance Testing

Cross-loadings and vis.

Check the bivariate correlations

```
round(cor(iris[c(1:4)]),2)
```

This gives us an indication of the correlations within the variates:



RStudio interface showing the iris dataset and the output of the canonical correlation analysis.

iris dataset (first 6 rows):

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa

Canonical Correlation Analysis Output:

```
> round(cor(iris[c(1:4)]),2)
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	1.00	-0.12	0.87	0.82
Sepal.Width	-0.12	1.00	-0.43	-0.37
Petal.Length	0.87	-0.43	1.00	0.96
Petal.Width	0.82	-0.37	0.96	1.00

CANONICAL CORRELATION

Data / Package Preparation

Bivariate Correlations

Canonical Correlations

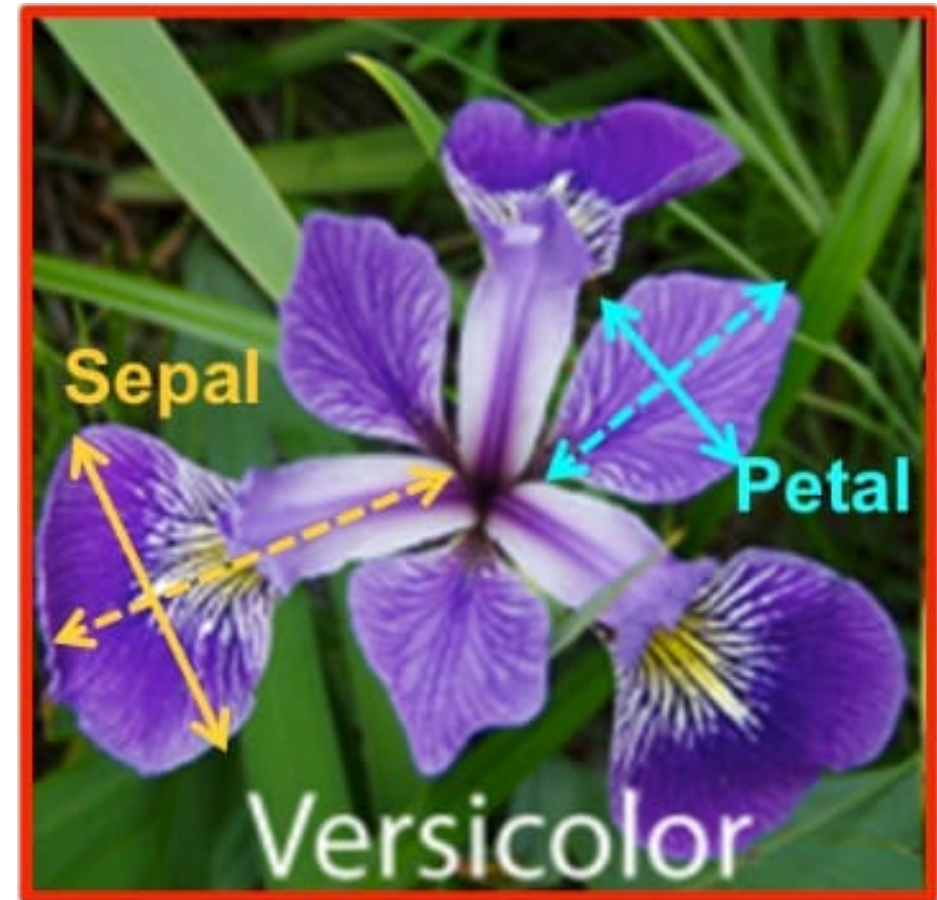
Significance Testing

Cross-loadings and vis.

Check the bivariate correlations

```
round(cor(iris[c(1:4)]),2)
```

This gives us an indication of the correlations within the variates:



CANONICAL CORRELATION

Data / Package Preparation

Bivariate Correlations

Canonical Correlations

Significance Testing

Cross-loadings and vis.

Define and create your canonical correlation model

```
cc1<-cca(sepal,petal)
```

Show the canonical correlation values for your model

```
summary(cc1)
```

The screenshot shows the RStudio interface. The top pane displays the 'iris' dataset with columns: Sepal.Length, Sepal.Width, Petal.Length, Petal.Width, and Species. The bottom pane shows the output of the `summary(cc1)` command, which includes the Canonical Correlation Analysis - Summary. The Canonical Correlations are highlighted in a blue box.

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa

Showing 1 to 7 of 150 entries, 5 total columns

```
> summary(cc1)
```

Canonical Correlation Analysis - Summary

Canonical Correlations:

CV 1	CV 2
0.9409690	0.1239369

Shared Variance on Each Canonical Variate:

CV 1	CV 2
0.88542265	0.01536035

CANONICAL CORRELATION

Data / Package Preparation

Bivariate Correlations

Canonical Correlations

Significance Testing

Cross-loadings and vis.

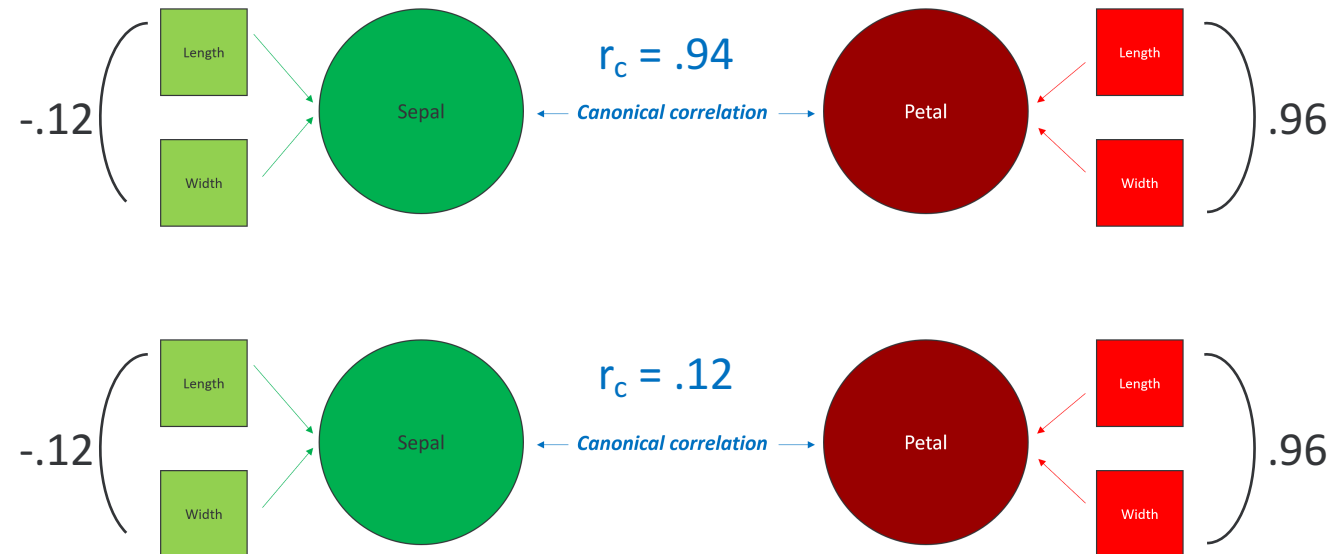
Define and create your canonical correlation model

```
cc1<-cca(sepal,petal)
```

Show the canonical correlation values for your model

```
summary(cc1)
```

* Note: scroll up (as this provides a lot of output)



RStudio interface showing the iris dataset and the output of the `summary(cc1)` command.

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa

Showing 1 to 7 of 150 entries, 5 total columns

Console Terminal Jobs

```
> summary(cc1)
```

Canonical Correlation Analysis - Summary

Canonical Correlations:

CV 1	CV 2
0.9409690	0.1239369

Shared Variance on Each Canonical Variate:

CV 1	CV 2
0.88542265	0.01536035

CANONICAL CORRELATION

Data / Package Preparation

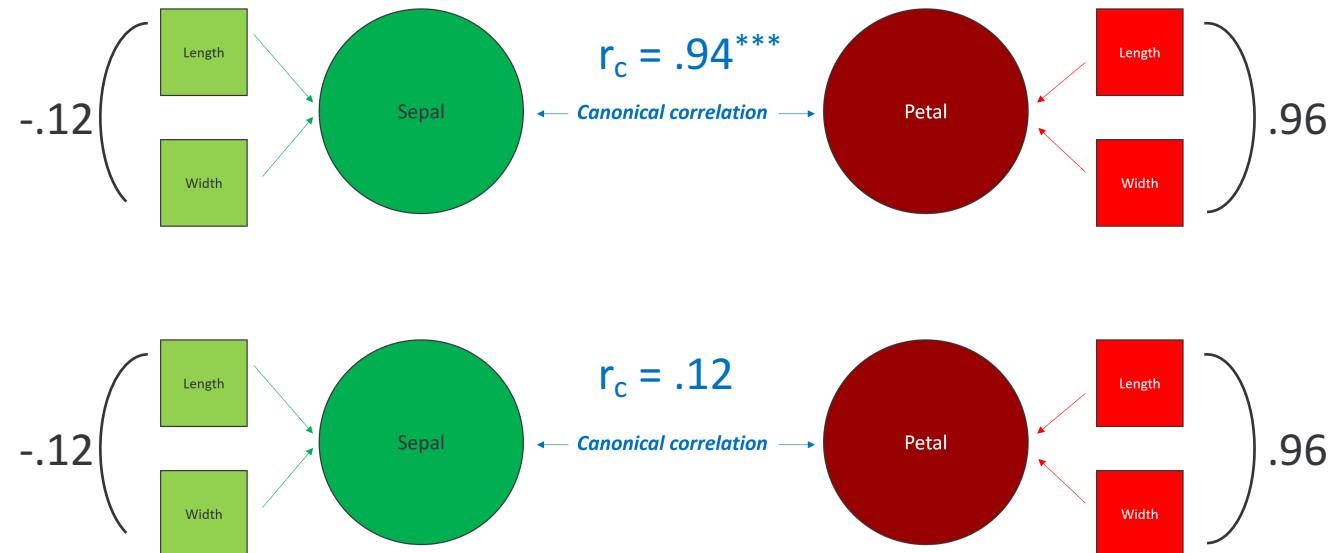
Bivariate Correlations

Canonical Correlations

Significance Testing

Cross-loadings and vis.

- Part of the output from the previous command gives us the chi-square tests
- Here only the first canonical correlation is significant, $\chi^2(4) = 319.66, p < .001$



The screenshot shows the RStudio interface with the following output:

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa

Showing 1 to 7 of 150 entries, 5 total columns

Console

```
CV 1      CV 2  
0.88542265 0.01536035
```

Bartlett's Chi-Squared Test:

	rho^2	Chisq	df	Pr(>X)
CV 1	0.88542	319.66076	4	<2e-16 ***
CV 2	0.01536	2.26775	1	0.1321

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

Canonical Variate Coefficients:

X Vars:

CANONICAL CORRELATION

Data / Package Preparation

Bivariate Correlations

Canonical Correlations

Significance Testing

Cross-loadings and vis.

We can also use other multivariate methods to assess for significance as well

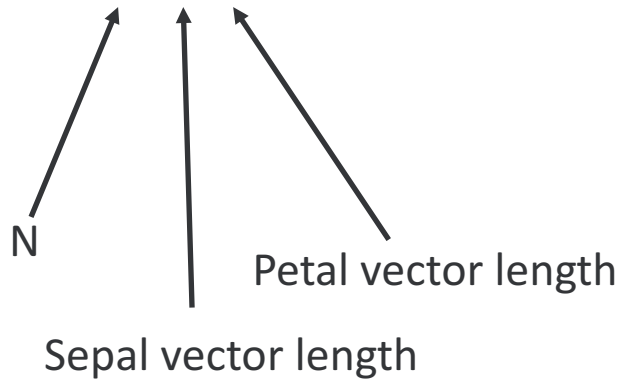
Install / Load the required packages

`install.packages("CCP")`

`library(CCP)`

Check Wilk's Lambda (or any other method)

`p.asym(cc1$corr,150,2,2,tstat = "Wilks")`



The screenshot shows the RStudio interface with the `iris` dataset loaded in the Environment pane. The Console pane displays the output of the `p.asym` function, which calculates Wilks' Lambda using an F-approximation (Rao's F).

```
> p.asym(cc1$corr,150,2,2,tstat = "wilks")
wilks' Lambda, using F-approximation (Rao's F):
      stat approx df1 df2 p.value
1 to 2: 0.1128174 144.337576 4 292 0.0000000
2 to 2: 0.9846396 2.293196 1 147 0.1320893
> p.asym(cc1$corr,150,2,2,tstat = "wilks")
+
> p.asym(cc1$corr,150,2,2,tstat = "wilks")
wilks' Lambda, using F-approximation (Rao's F):
      stat approx df1 df2 p.value
1 to 2: 0.1128174 144.337576 4 292 0.0000000
2 to 2: 0.9846396 2.293196 1 147 0.1320893
> |
```

CANONICAL CORRELATION

Data / Package Preparation

Bivariate Correlations

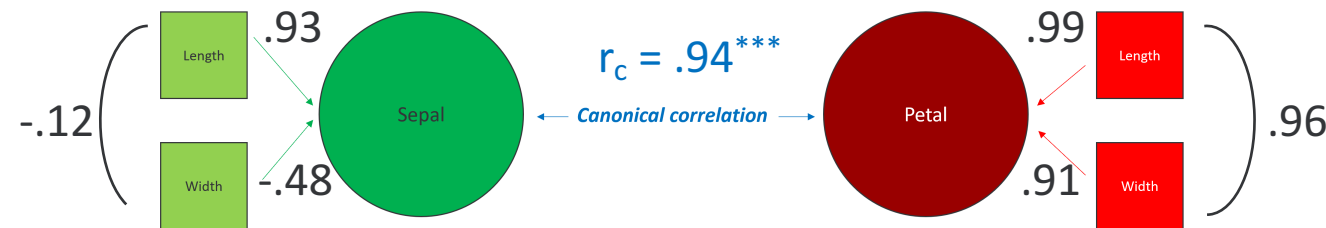
Canonical Correlations

Significance Testing

Cross-loadings and vis.

The yacca package and the cca function we ran earlier provides us with the structural loadings (note: we'll only comment on the first one as the second was not significant)

You can re-run if you lost the previous output
`summary(cc1)`



```
iris
Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1 5.1 3.5 1.4 0.2 setosa
2 4.9 3.0 1.4 0.2 setosa
...
Showing 1 to 2 of 150 entries, 5 total columns

Console Terminal Jobs
~/
Sepal.Length 1.000000 0.370077
Sepal.Width -0.854991 2.146275

Y Coefficients:
              CV 1      CV 2
Petal.Length 0.8491249 -1.918708
Petal.Width -0.6938022  4.809531

Structural Correlations (Loadings) - X Vars:
              CV 1      CV 2
Sepal.Length 0.9290009 0.3700774
Sepal.Width -0.4767332 0.8790480

Structural Correlations (Loadings) - Y Vars:
              CV 1      CV 2
Petal.Length 0.9897548 0.1427778
Petal.Width 0.9144533 0.4046915

Aggregate Redundancy Coefficients (Total Variance Explained):
X | Y: 0.4896823
Y | X: 0.805307
```

CANONICAL CORRELATION

Data / Package Preparation

Bivariate Correlations

Canonical Correlations

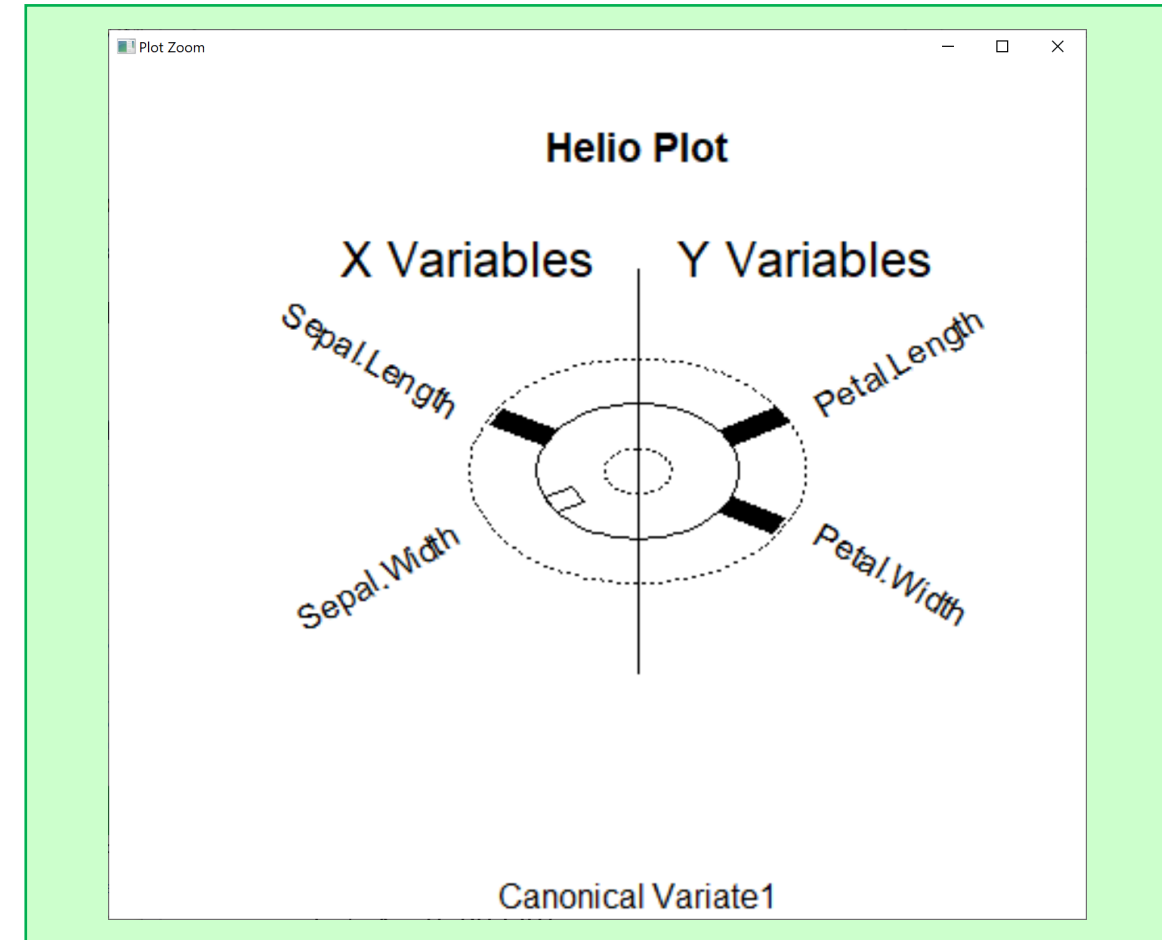
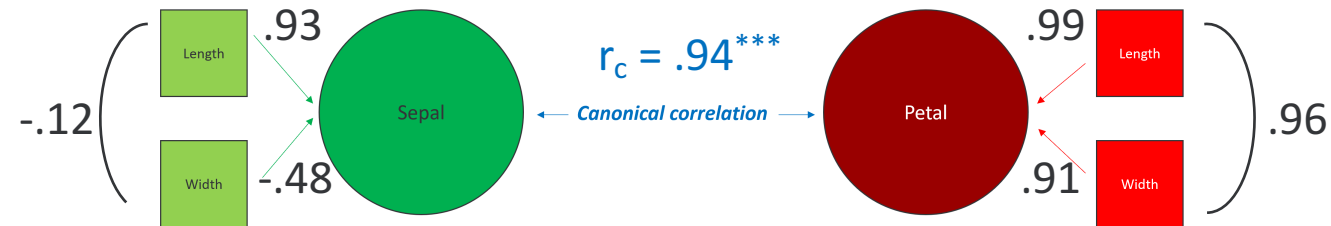
Significance Testing

Cross-loadings and vis.

The yacca package also allows us to create helio plots to visualize our data

```
# Create helio plot  
Helio.plot(cc1)
```

Comparing this plot to our model below, what do you think the black bars represent? What about the white?



CANONICAL CORRELATION

Data / Package Preparation

Bivariate Correlations

Canonical Correlations

Significance Testing

Cross-loadings and vis.

Now load the [LifeCycleSavings](#) data file (this is a base R file) and repeat what we just did
(**Hint:** inspect the data frame first, and think about which variables will go to which sets)

About the data:

Under the life-cycle savings hypothesis as developed by Franco Modigliani, the savings ratio (aggregate personal saving divided by disposable income) is explained by per-capita disposable income, the percentage rate of change in per-capita disposable income, and two demographic variables: the percentage of population less than 15 years old and the percentage of the population over 75 years old. The data are averaged over the decade 1960–1970 to remove the business cycle or other short-term fluctuations.