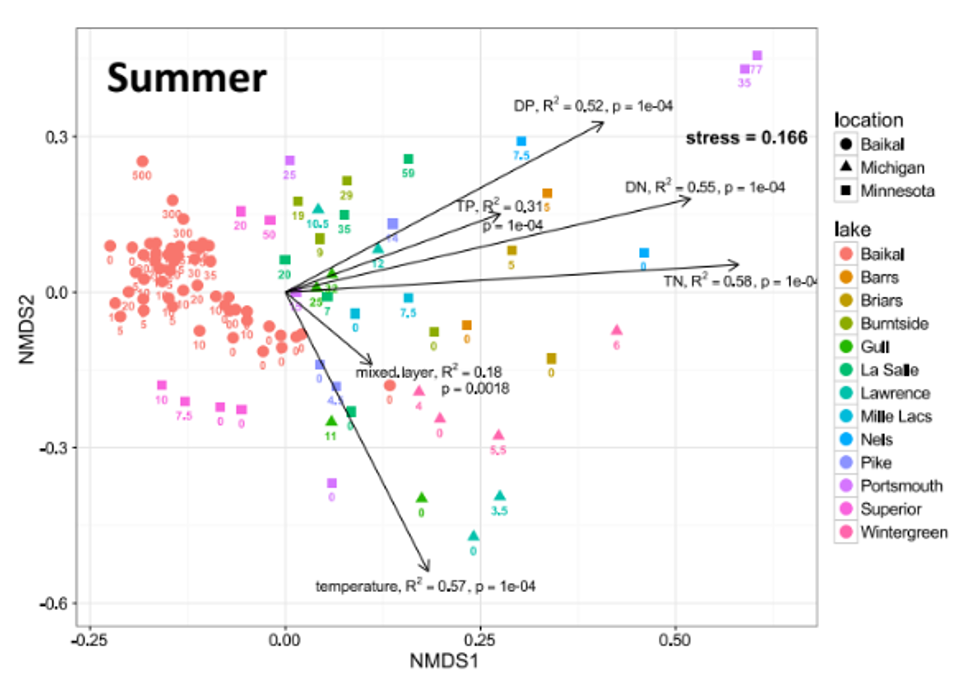
Lecture 16 - Multivariate Statistics

Bill Perry

# Introduction to Multivariate Statistics

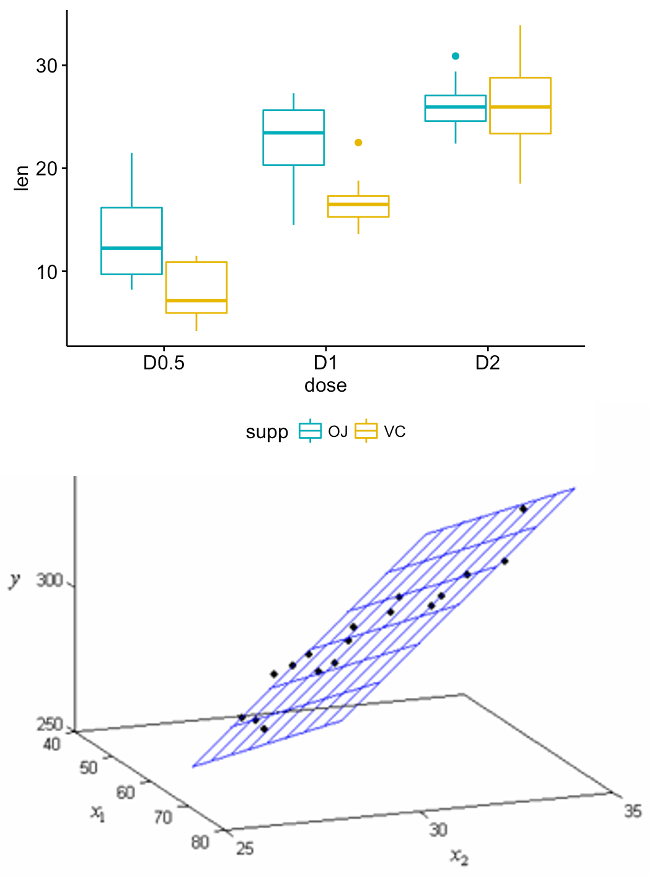
## Overview

* Multivariate data: multiple variables per object
* Types of multivariate analyses
  + Functional vs. structural methods
  + R-mode vs. Q-mode analyses
* Eigenvectors, eigenvalues, and components
* Distance and dissimilarity measures
* Data transformations and standardization
* Screening multivariate data
* MANOVA



# Multivariate Data Structure

* Multiple variables recorded about each object (individual, quadrat, site, etc.)
  + or responses that are from the same treatment factor
  + length, weight, width, color, spines, etc
* Objects: rows (i = 1 to n)
* Variables: columns (j = 1 to p)
* Examples:
  + Stream sites with multiple chemical parameters
  + Species with multiple morphological traits
  + Sample units with multiple species abundances



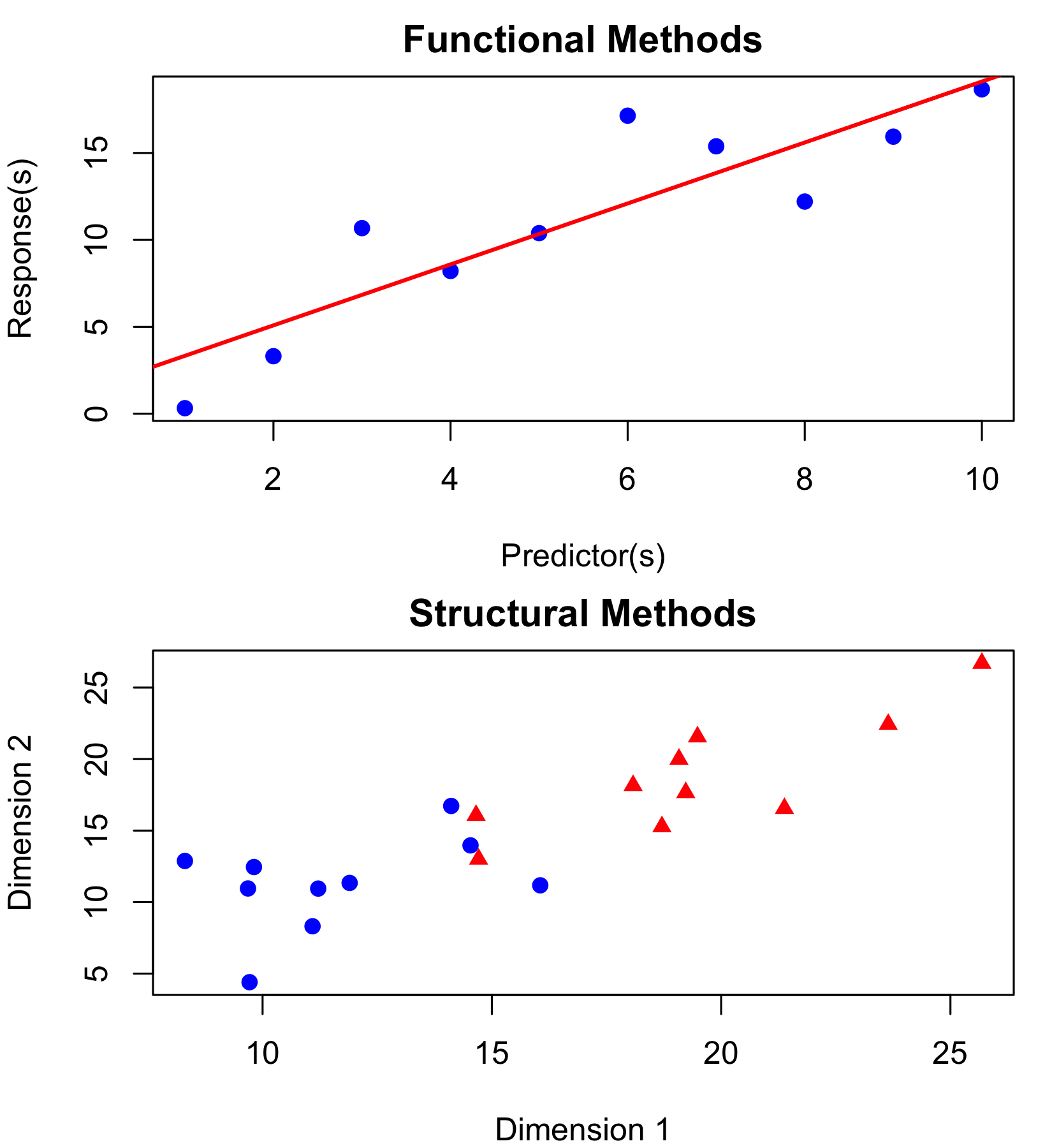
|  |
| --- |
| Multivariate data vs. multivariate analysis |
| We’ve already seen multivariate data in multiple regression and multi-factor ANOVA  **Now we’ll look at cases with multiple response variables.** |

# Functional vs. Structural Methods

## Functional vs. Structural Methods

**Functional methods**: - Clear response and predictor variables - Goal: relate Y’s to X’s - Examples: MANOVA, PERMANOVA

**Structural methods**: - Find patterns/structure in data - Often no clear predictors - Examples: PCA, NMDS, Cluster Analysis



# Functional Methods Examples

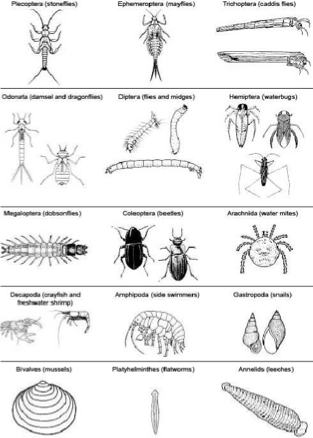
Example 1:

* sample 30 stream sites (objects)
* record TP, TN, pH, DO, chloride concentration, etc.
* each parameter is a variable

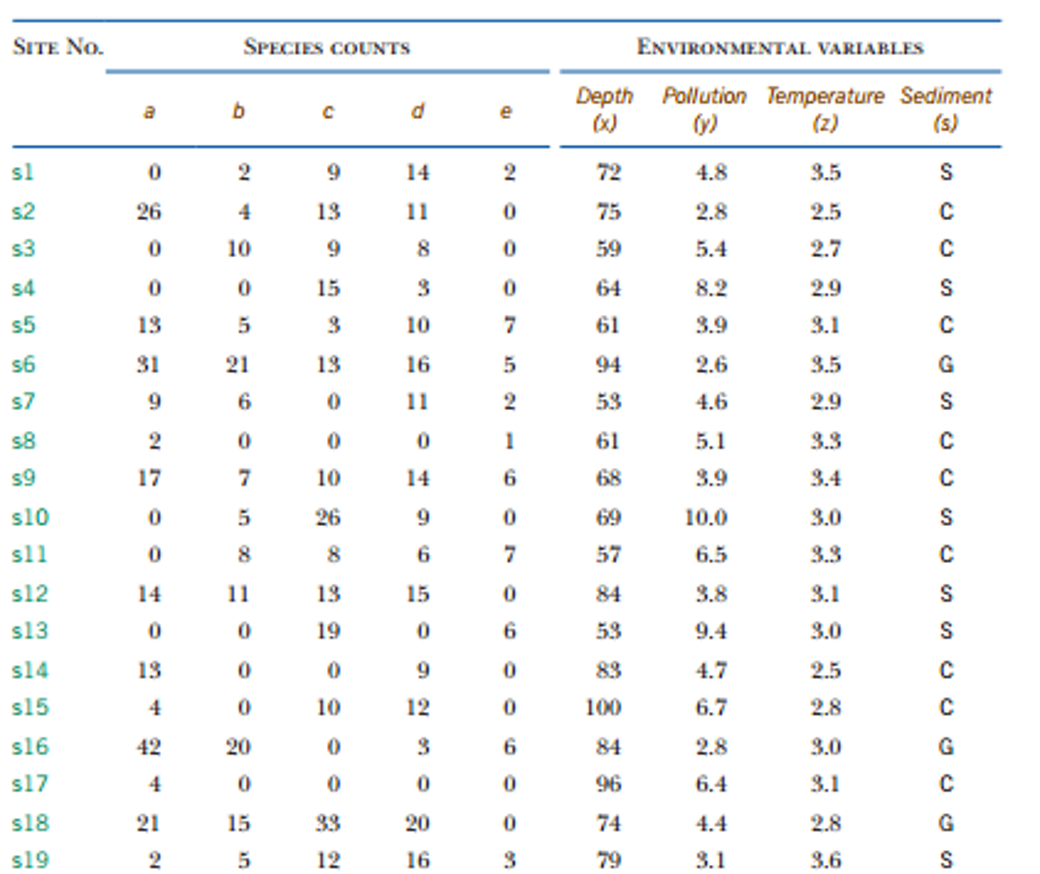
Example 2:

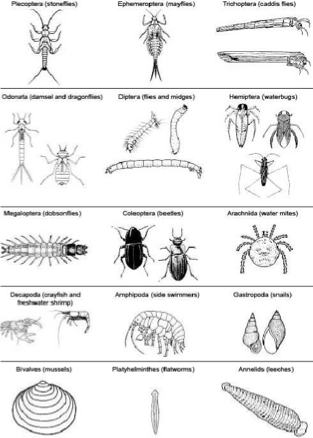
* sample 30 stream sites (objects)
* collect benthic invertebrates
* each species is now a variable

Sometimes combine both…



# Functional Methods Visualization

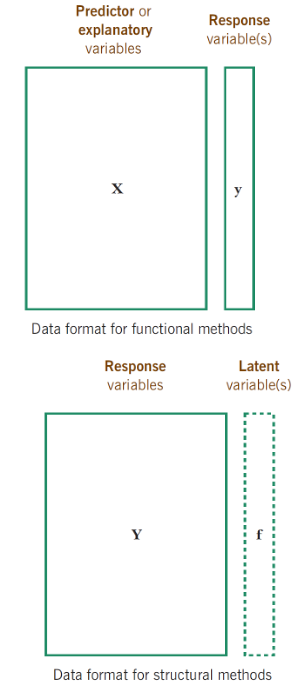




# Ecological Multivariate Methods Overview

Can divide ecological MV methods into “functional” and “structural”

* Functional methods: clear response variable(s) and predictor variables. Goal is to relate Ys to Xs (regression, MANOVA, ANOSIM, PERMANOVA).
* Structural methods: concerned with finding structure /pattern in the data. Often no clear predictor variables (PCA, NMDS, Cluster analysis).

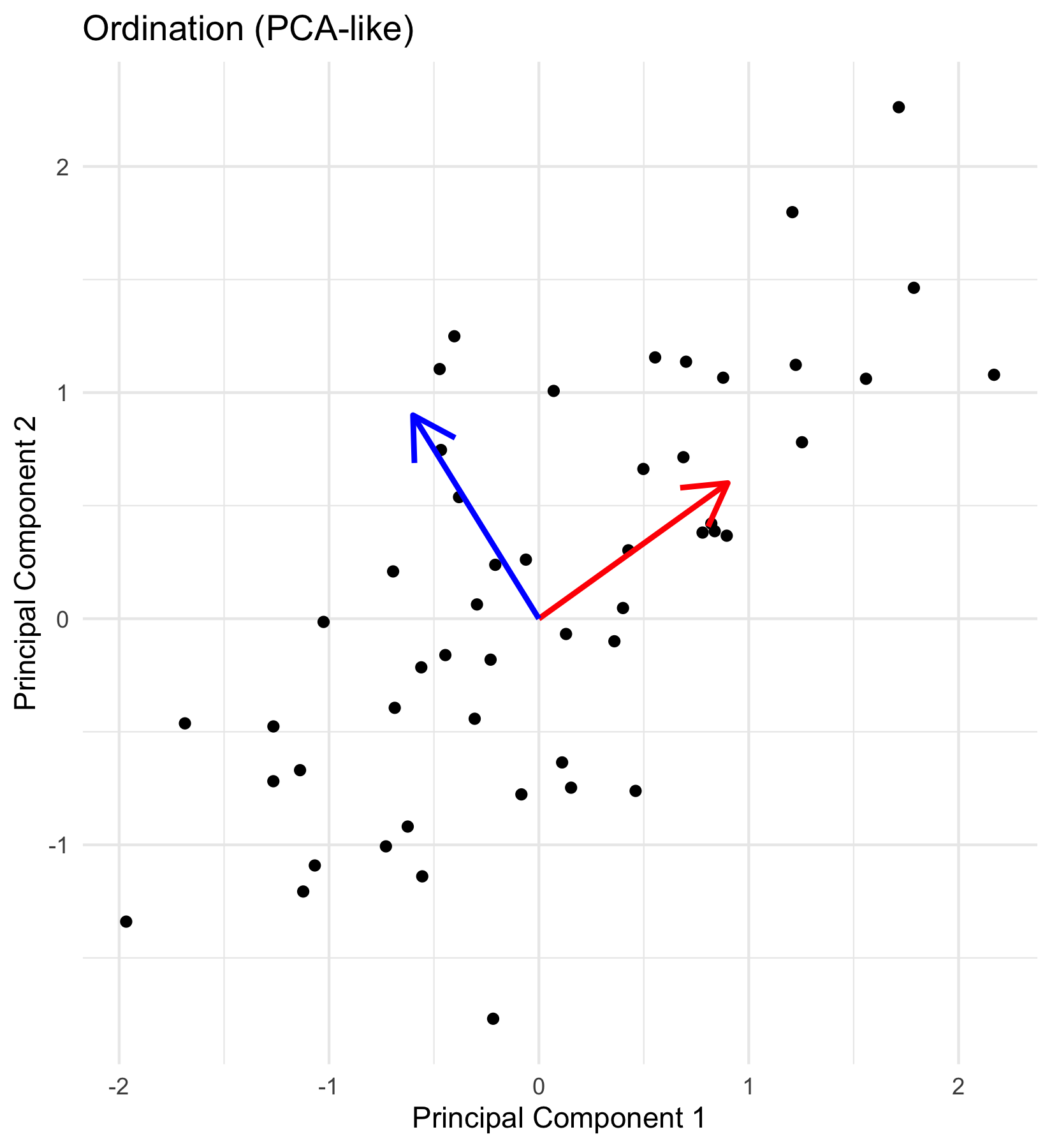


# Structural Methods: Two Approaches

## Two Main Approaches

**Scaling/Ordination Methods**: - Reduce dimensions with new derived variables - Summarize patterns in data - Examples: PCA, CCA

**Dissimilarity-Based Methods**: - Measure dissimilarity between objects - Visualize relationships between objects - Examples: NMDS, Cluster Analysis



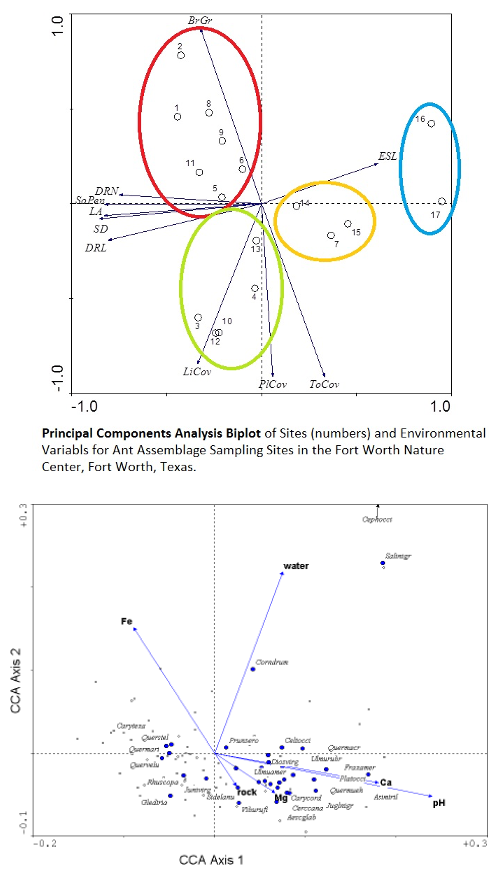
# Structural Methods: Scaling/Ordination

Structural methods can be divided further into:

Methods based on scaling or ordination

Goal: reduce number of vars by deriving new variables that summarize data.

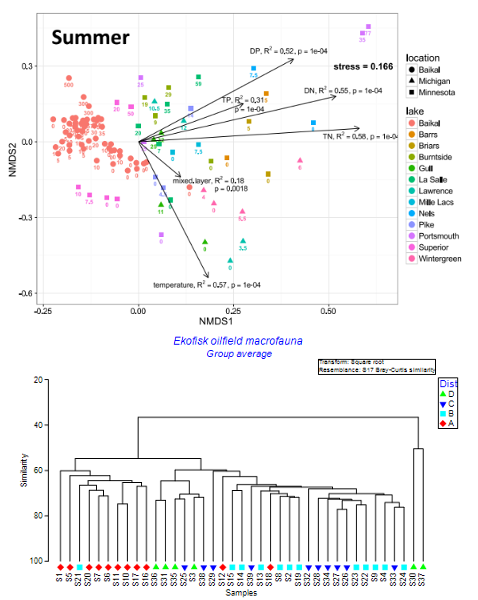
Examples include PCA, CCA



# Structural Methods: Dissimilarity-Based

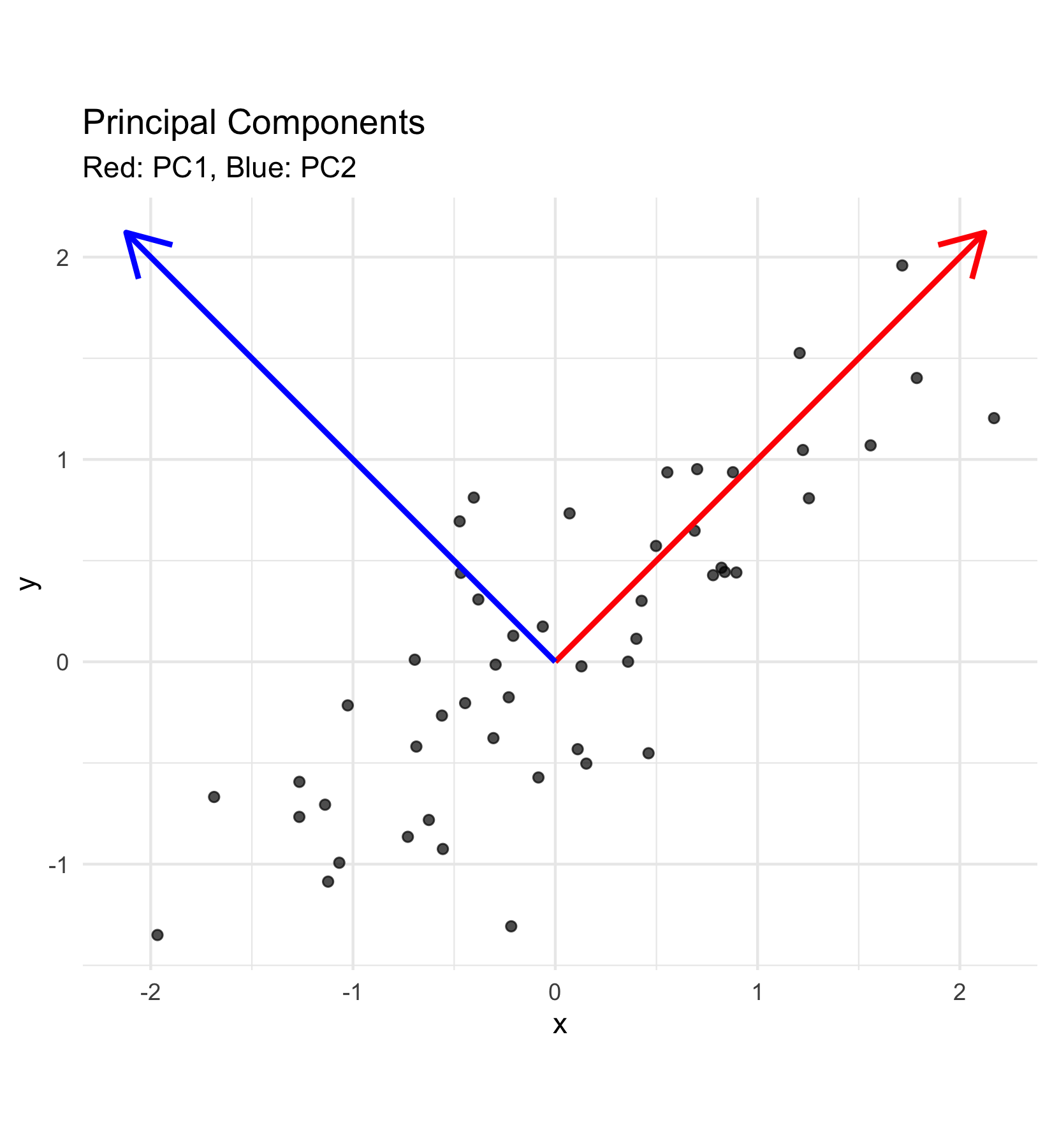
Structural methods can be divided further into:

* Methods based on dissimilarity measurements
* Goal: measure and graphically show degree of dissimilarity between objects.
* Examples include (N)MDS and cluster analysis



# Eigenvectors, Eigenvalues, and Components: Concept

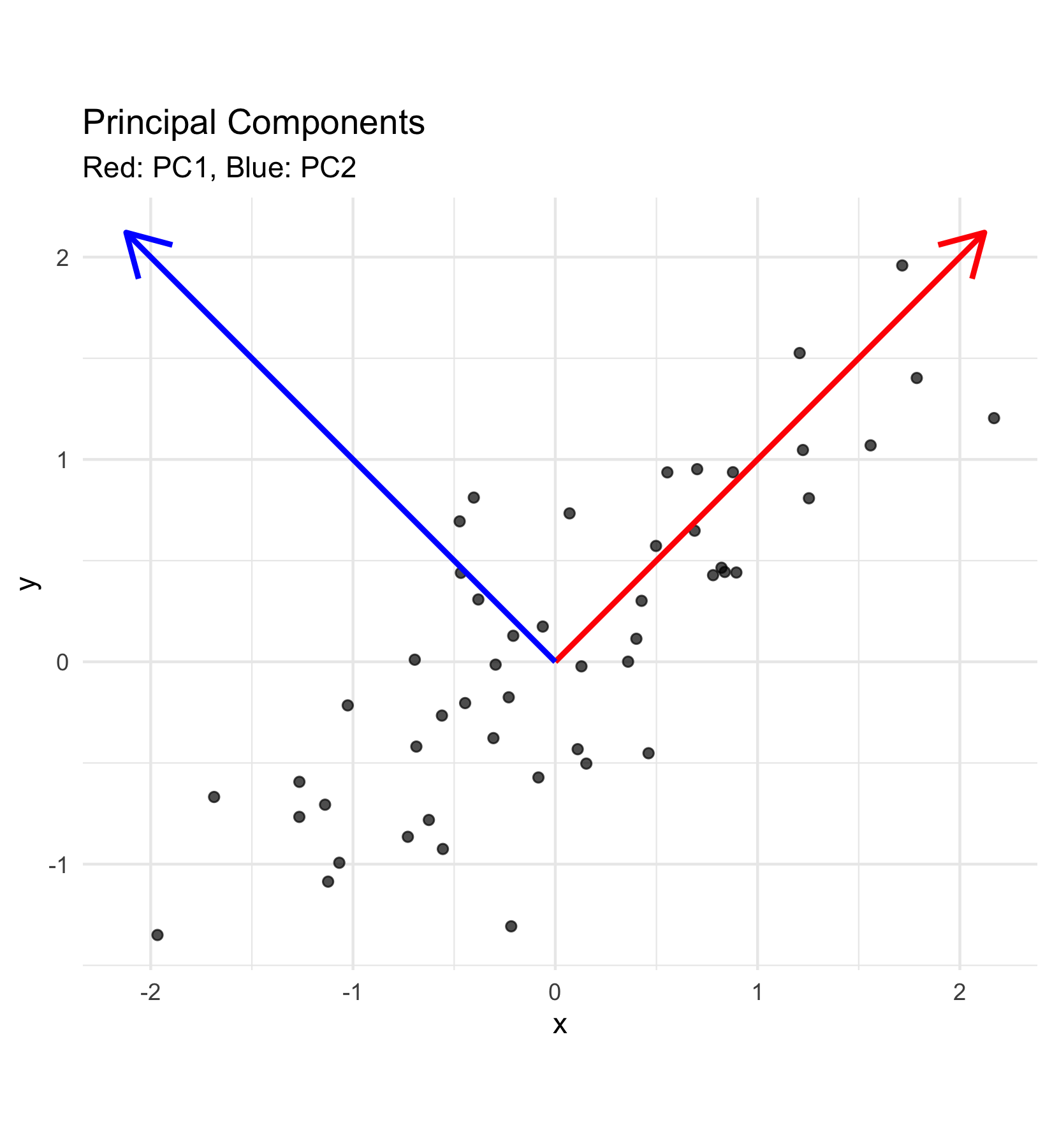
* Goal: derive new variables (principal components) that explain variation in data
* Components are linear combinations of original variables:
  + zik = c1yi1 + c2yi2 + … + cpyip
* Properties of derived variables:
  + First component explains most variation
  + Second explains most remaining variation
  + Components are uncorrelated with each other
  + As many components as original variables



# Eigenvectors and Components: Interpretation

How to think about the new values

* zik is value of new variable k for object I
* yi1- yip are values of original variables for object i
* c1-cp are coefficients that show importance of the original variables to new derived variable

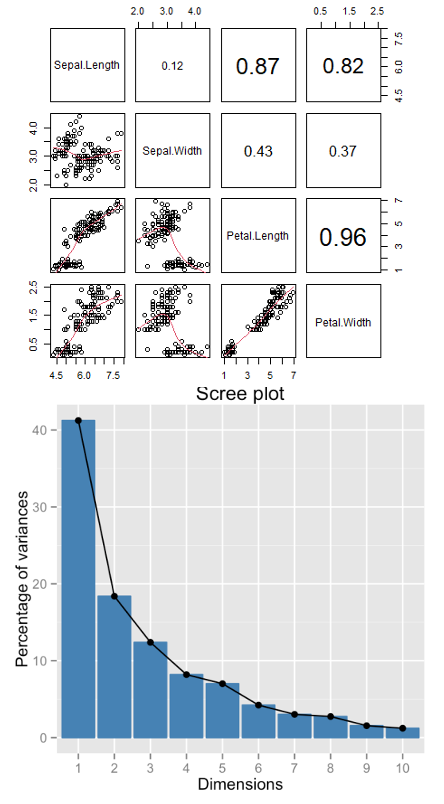


|  |
| --- |
| Key concept |
| Eigenvalues (λ) represent the amount of variation explained by each new derived variable, while eigenvectors contain the coefficients showing how original variables contribute to each component. |

# Eigenvalues and Components: Properties

Derived variables are found so that:

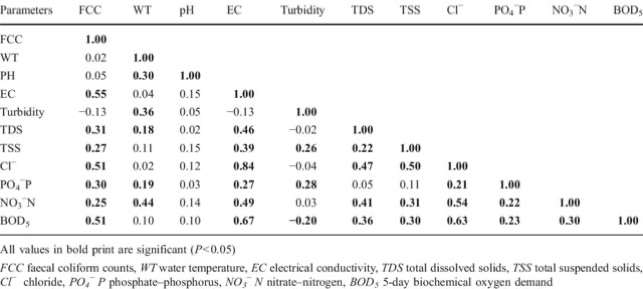
* First derived variable explains most of the variation in the data
* Second most of the remaining variation
* And so on…
* As many derived variables as original variables (p)
* Derived variables are uncorrelated with each other



# Eigenvalues and Eigenvectors: Mathematical Details

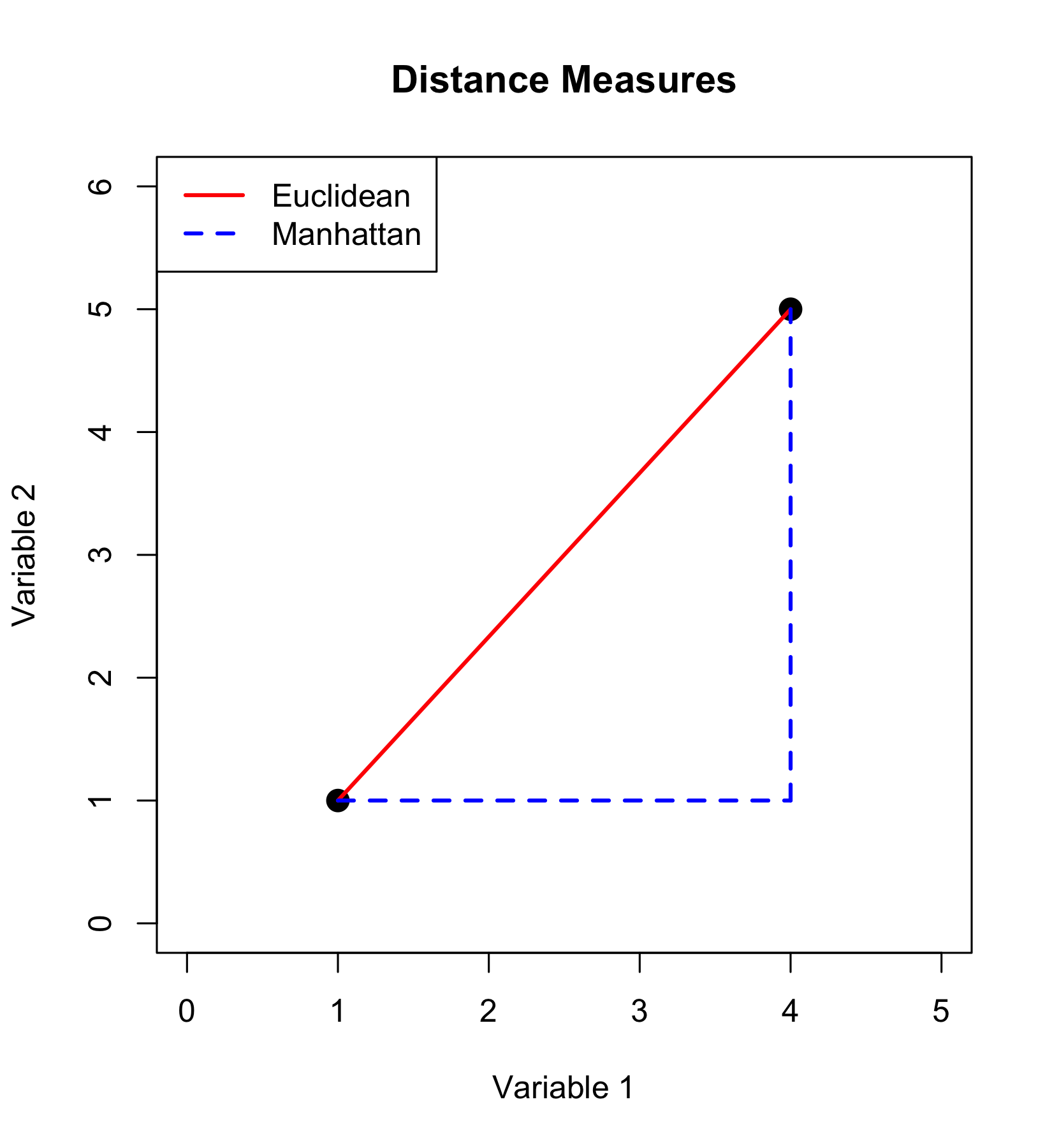
* Eigenvalues (latent roots) represent amount of variation in data explained by the new k= 1 to p derived variables (λ1, λ2 …λp).
* Eigenvalues are population parameters and are estimated using ML to get sample statistics (l1, l2…lp)
* Eigenvectors are lists of coefficients (c) that show contribution of original variables to new, derived variables
* Each new variable has an eigenvalue and an eigenvector
* New variables (components) are derived from a p x p covariance or correlation matrix of original variables

# Eigenvalue Matrix Representation



# Distance and Dissimilarity Measures: Concept

* Measure how different objects are in multivariate space
* Common measures:
  + **Euclidean distance**: direct geometric distance
  + **Manhattan distance**: sum of absolute differences
  + **Bray-Curtis**: good for species abundance data
  + **Kulczynski**: for abundance data with zeros
* Used in cluster analysis, MDS, and other techniques
* Create dissimilarity matrices for analysis



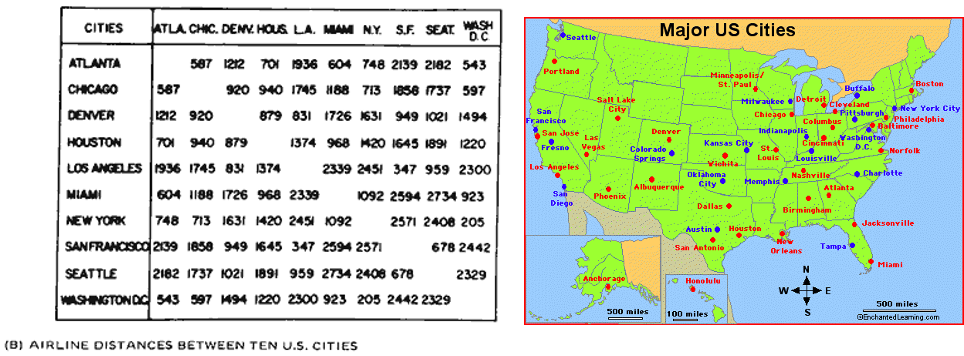
# Distance and Dissimilarity: Background

* Previous approach relies on analysis of covariance/ correlation bw variables
* Another class of MV analysis uses measures of similarity/dissimilarity bw objects (MDS, cluster analysis)
* Similarity/dissimilarity indices measure how alike/different objects (e.g. Lakes) are in MV space
* Many measures of dissimilarity (Euclidean, Manhattan, Bray-Curtis, etc, etc)



# Dissimilarity Matrix Representation

Dissimilarity is often represented as a dissimilarity matrix



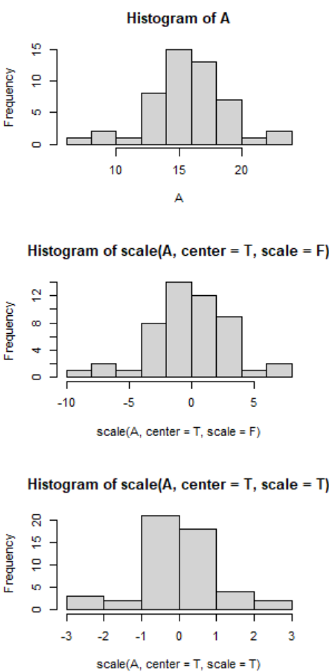
# Data Transformations: Common Approaches

Data transformation is common and useful in MV analyses

* Log transformation is common in PCA/CCA analyses based on eigenvectors, since linearizing relationships bw variables will improve extraction of eigenvectors
* Fourth root transform is very common and sometimes “blanket recommended” for analysis of species composition data (each variable is a species w counts- MDS, cluster analysis). Idea is to lessen importance of common and abundant species

# Data Standardization: Methods

* Data standardization is also common; adjusts data so all variables have same means and/or variance
  + Centering- mean subtracted from each value (new mean=0)
  + Standardization- centered observations divided by SD (mean=0, sd=1)
* Crucial for analyses of variables measured in different units
* More ambiguous for species abundance data

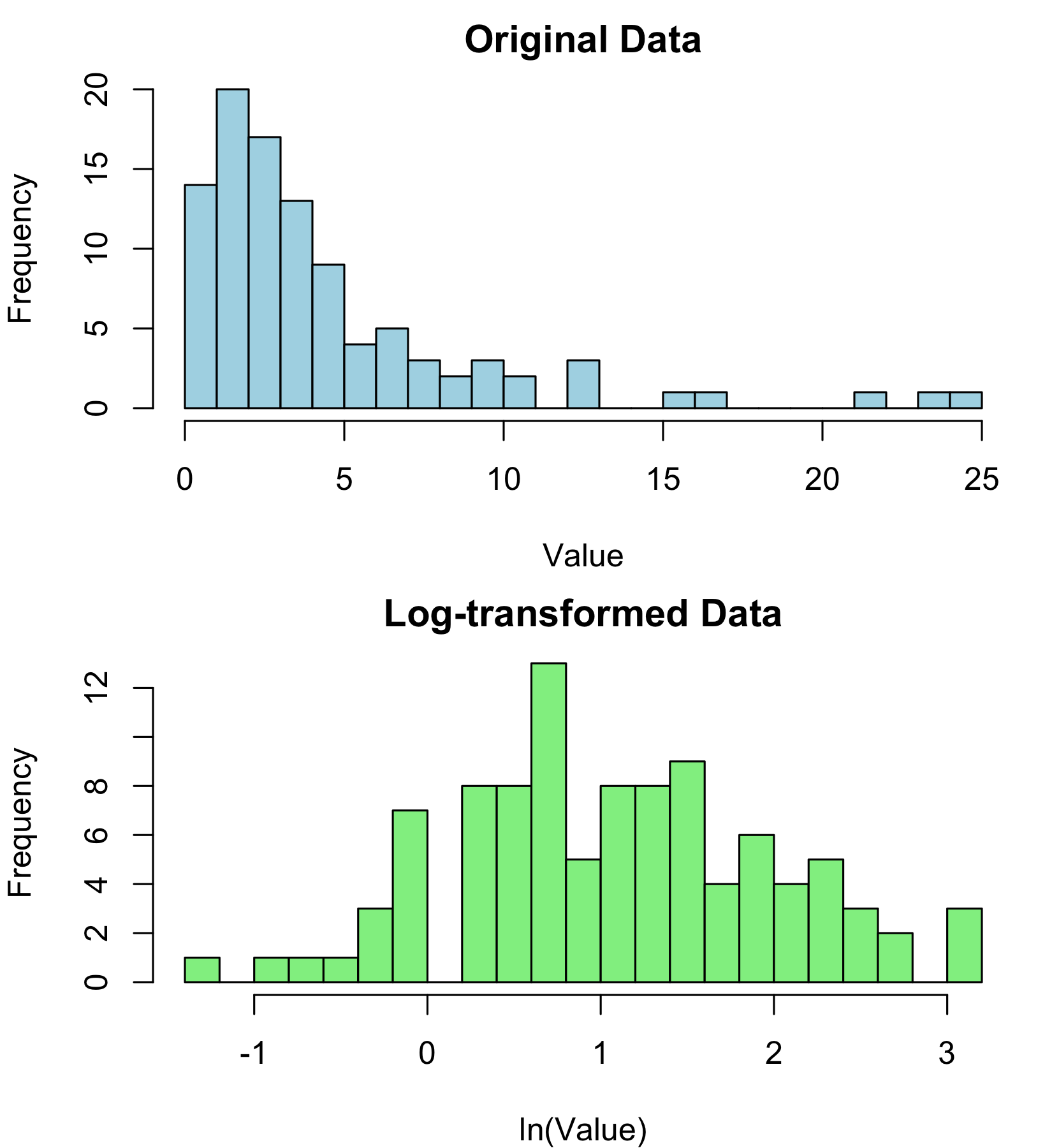


# Data Transformations & Standardization: Visual

## Common Approaches

**Transformations**: - Log transformation for skewed data - Root transformations for count data  
- Fourth-root for species abundance data

**Standardization**: - Centering: subtract mean (mean = 0) - Standardization: divide by SD (SD = 1) - Crucial for variables with different units - May not be appropriate for species data

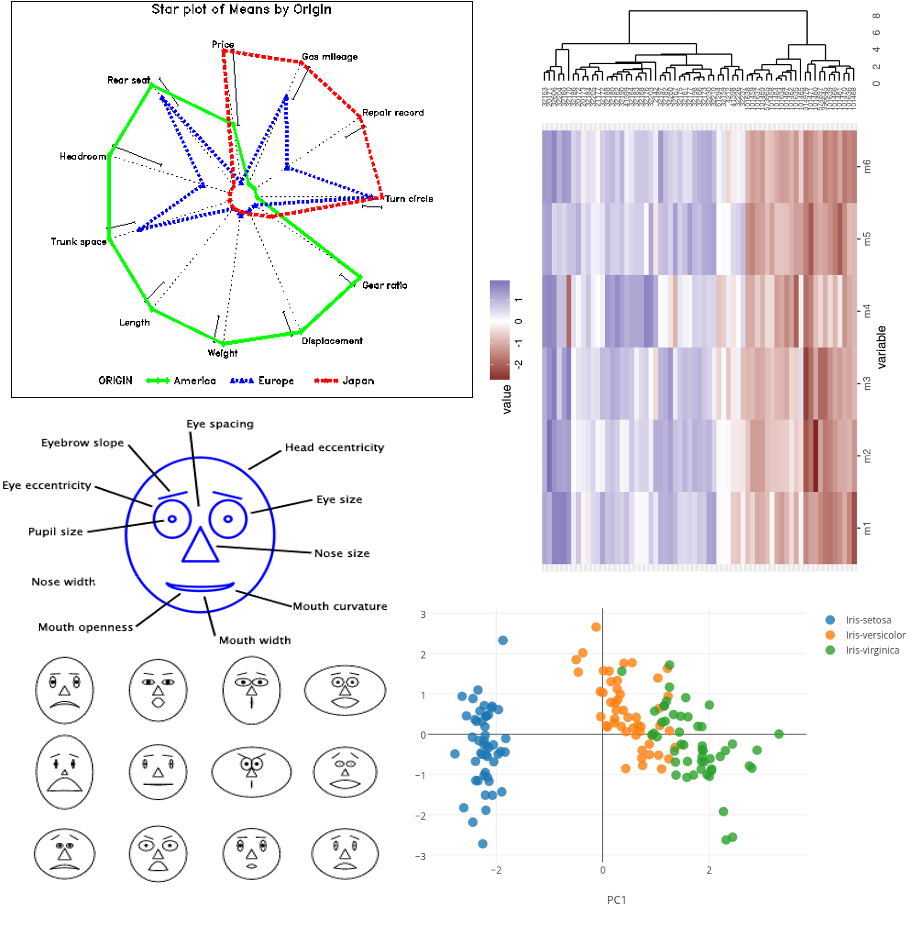


|  |
| --- |
| Why standardize? |
| Standardization ensures all variables contribute equally to the analysis regardless of their original units or scales of measurement. Without it, variables with larger values or variances would dominate the results. |

# Multivariate Graphics Options

## Visual Representation Methods

* **SPLOMS/Scatterplot Matrices**: show bivariate relationships
* **Star plots**: display multiple variables per object
* **Chernoff faces**: represent variables as facial features
* **Heatmaps**: visualize data matrices with color
* **Biplots**: show objects and variables together
* **Ordination plots**: visualize relationships in reduced dimensions

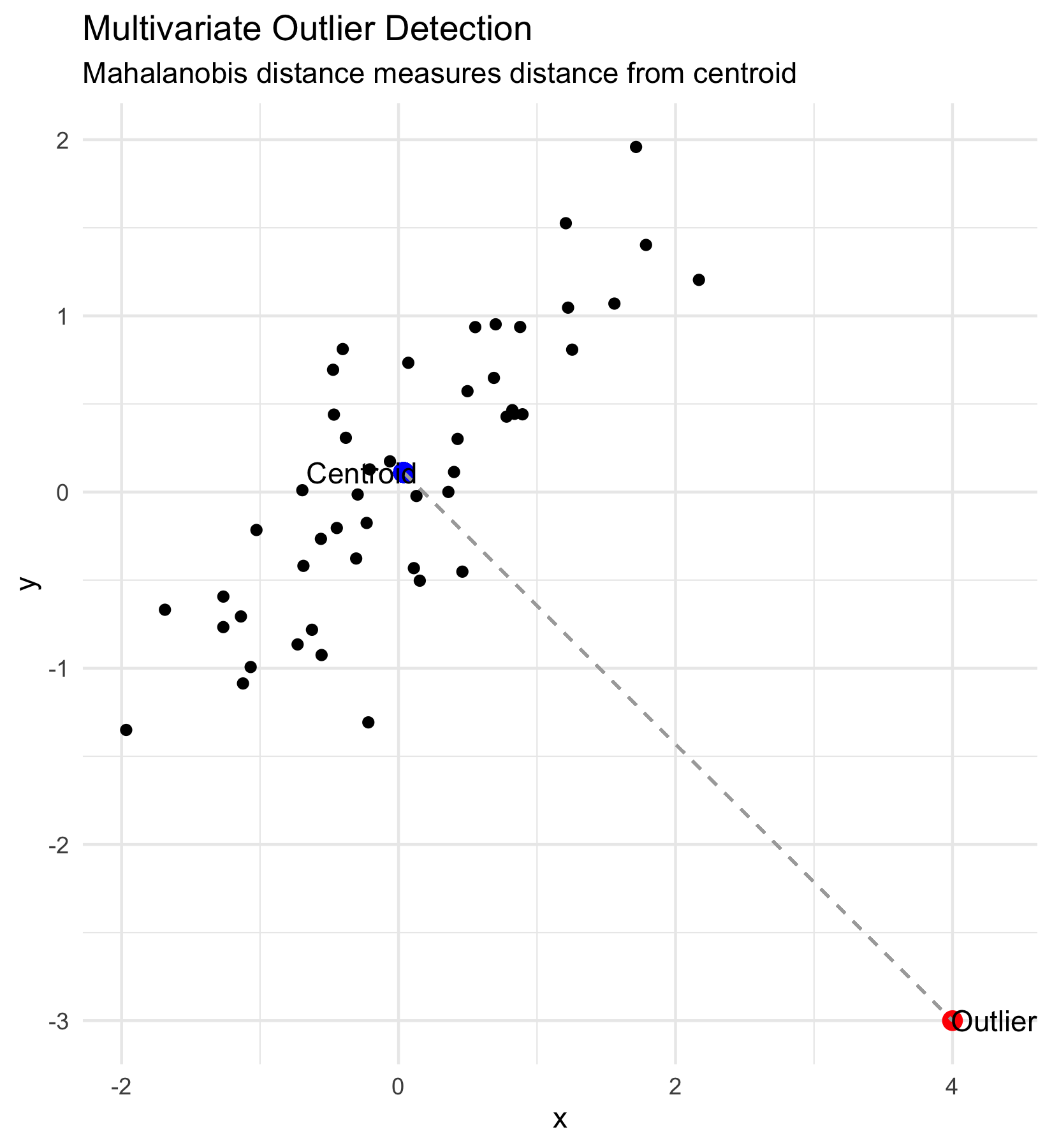


# Screening Multivariate Data: Outliers and Missing Data

## Key Issues to Check

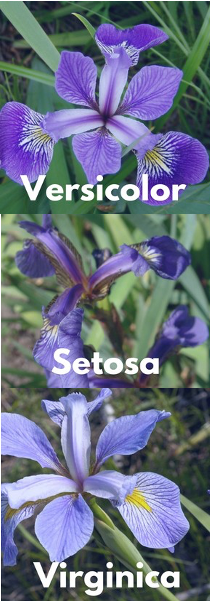
**Multivariate Outliers**: - Objects with unusual patterns across variables - Detected with Mahalanobis distance (d²) - Test against χ² distribution with p df

**Missing Observations**: - Common approaches: - Deletion: remove affected object or variable - Imputation: estimate missing values - Maximum likelihood methods - Multiple imputation



# MANOVA: Introduction

* Multivariate extension of ANOVA
* Tests for differences in group centroids based on multiple response variables
* Advantages over multiple ANOVAs:
  + Controls family-wise error rate
  + Accounts for correlations between variables
  + More powerful when variables are correlated
* Common test statistics:
  + Wilk’s lambda (λ)
  + Pillai’s trace
  + Hotelling-Lawley trace
* Famous dataframe built into R is the iris dataset



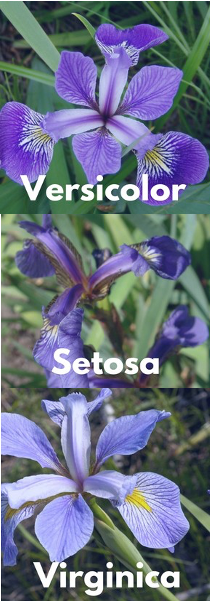
# MANOVA: Iris Dataset Example

Morphometric measurements on n=150 flowers

Response vars: Septal length + width, petal length + width

Predictor variable: species

Question: are there differences bw species?



# MANOVA: Data Structure

Morphometric measurements on n=150 flowers

Response vars: Septal length + width, petal length + width

Predictor variable: species

Question: are there differences bw species?

iris\_df <- iris %>% clean\_names()  
iris\_long\_df <- iris\_df %>% pivot\_longer(cols = -species,   
 names\_to = "variable",   
 values\_to = "measure")  
write\_csv(iris\_df, "data/iris.csv")  
head(iris\_df)

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

# MANOVA: Data Visualization

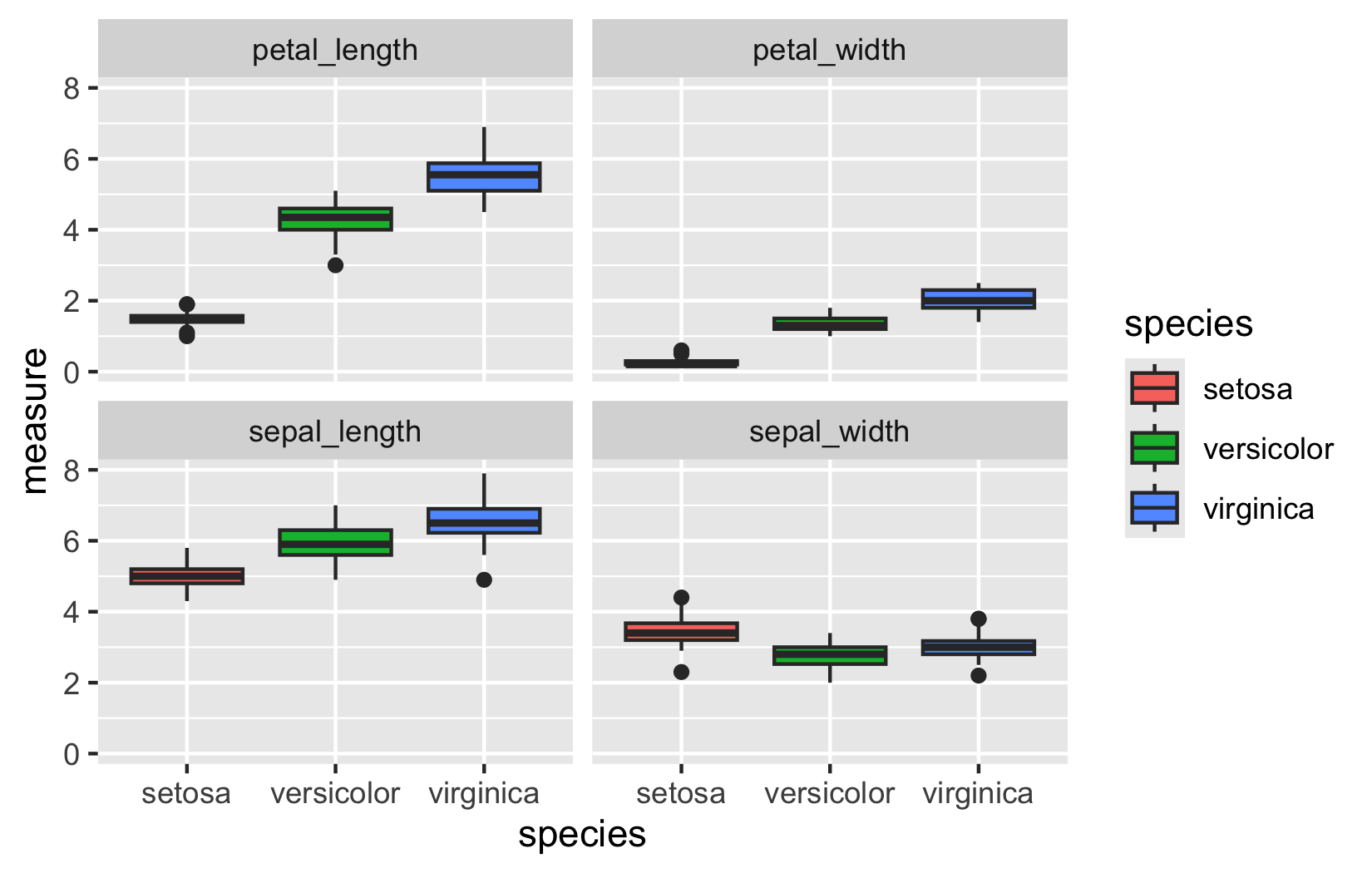
Morphometric measurements on n=150 flowers

Response vars: Septal length + width, petal length + width

Predictor variable: species

Question: are there differences bw species?

iris\_long\_df %>% ggplot(aes(species, measure, fill=species))+  
 geom\_boxplot()+  
 facet\_wrap(.~variable)



# MANOVA vs. Multiple ANOVAs

One approach:

series of 1-way ANOVAs

for example —>

But:

* Variables and tests are not independent
* Multiple testing problem can reduce power
* MANOVA considers all response variables simultaneously

sepal\_model <- aov(sepal\_length~species, data = iris\_df)  
Anova(sepal\_model, type = 3)

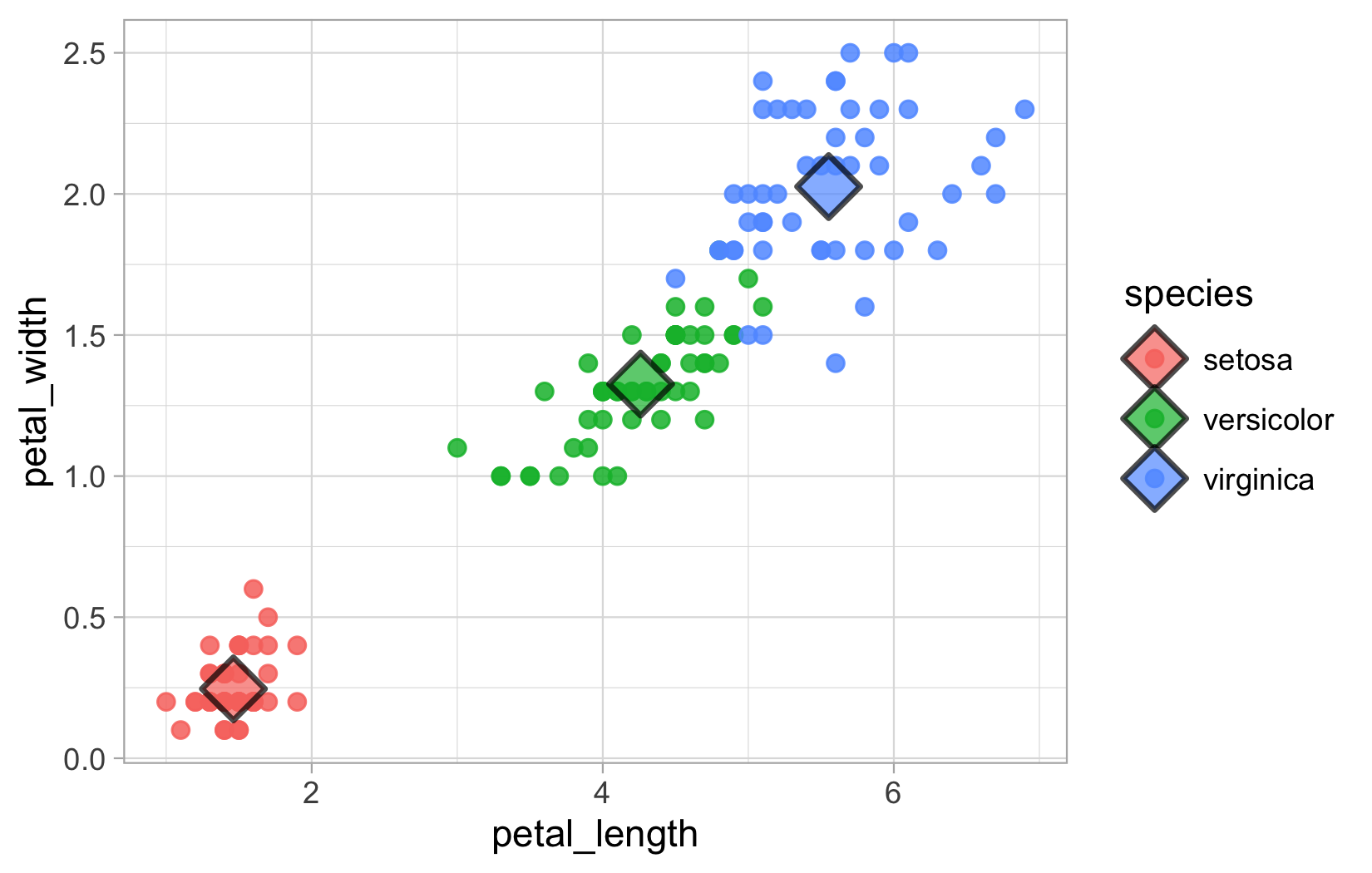
Anova Table (Type III tests)  
  
Response: sepal\_length  
 Sum Sq Df F value Pr(>F)   
(Intercept) 1253.00 1 4728.16 < 2.2e-16 \*\*\*  
species 63.21 2 119.26 < 2.2e-16 \*\*\*  
Residuals 38.96 147   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# MANOVA: Centroids vs. Means

Instead of means compare centroids

but for all of the variables not just two

mean\_points <- iris\_df %>%   
 group\_by(species) %>%   
 summarise(mean\_length = mean(petal\_length),   
 mean\_width = mean(petal\_width),  
 .groups = 'drop')  
  
iris\_plot<-iris\_df %>%   
 ggplot(aes(x=petal\_length, y=petal\_width, color=species)) +  
 geom\_point(alpha = 0.85, size = 2) +  
 geom\_point(data=mean\_points,   
 aes(x=mean\_length, y=mean\_width, fill=species),   
 shape=23, color="black", stroke=1.2,alpha = .7,  
 size=6) +  
 theme\_light()  
iris\_plot

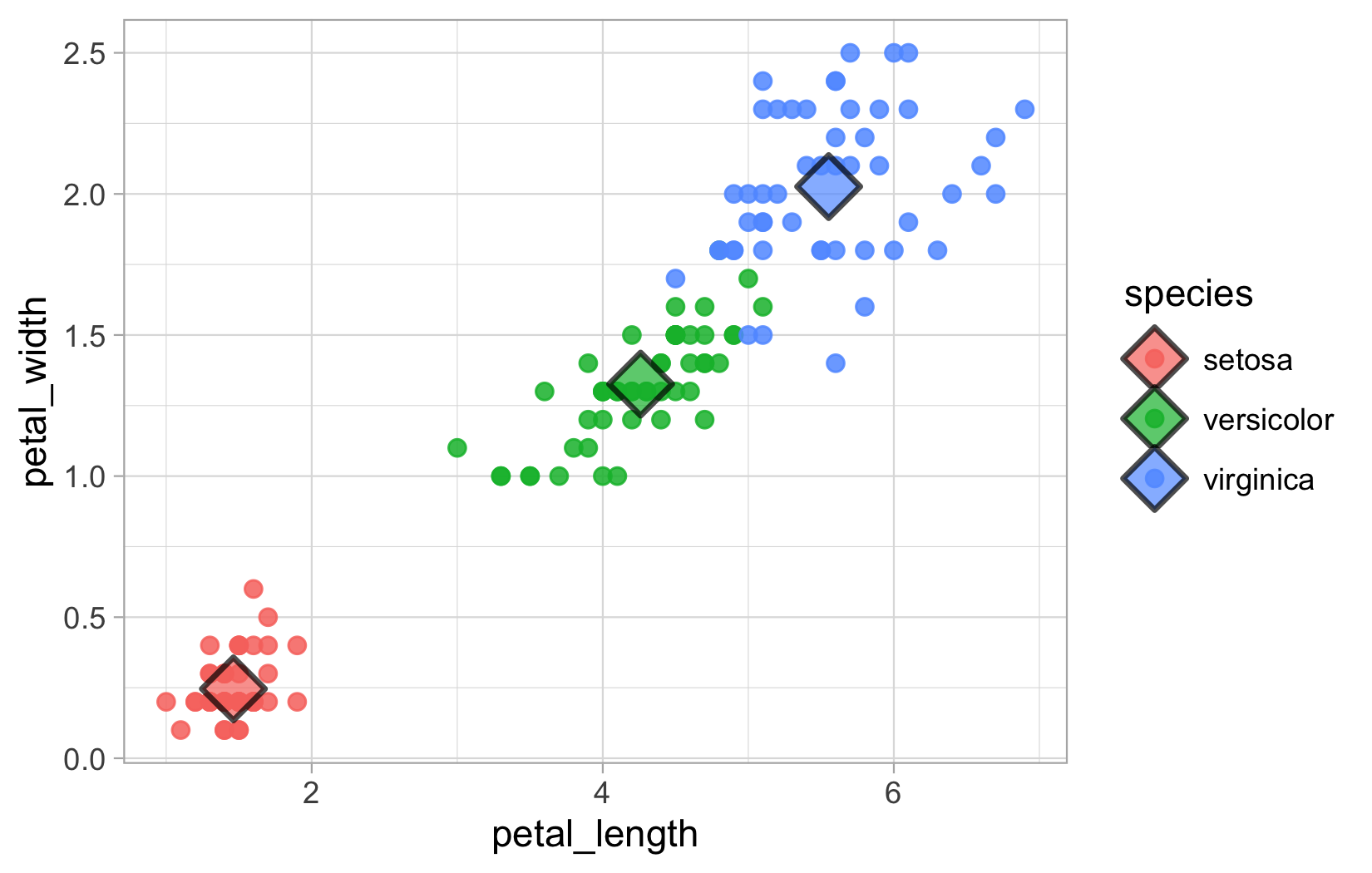


# MANOVA: SSCP Matrices

A one-way MANOVA tests the Ho that there are no differences in population centroids

* Ho tested by partitioning variance, but instead of SS, use SSCP matrices:
* H matrix: between group SSCP
* E matrix: within group SSCP
* T matrix: total SSCP

iris\_plot

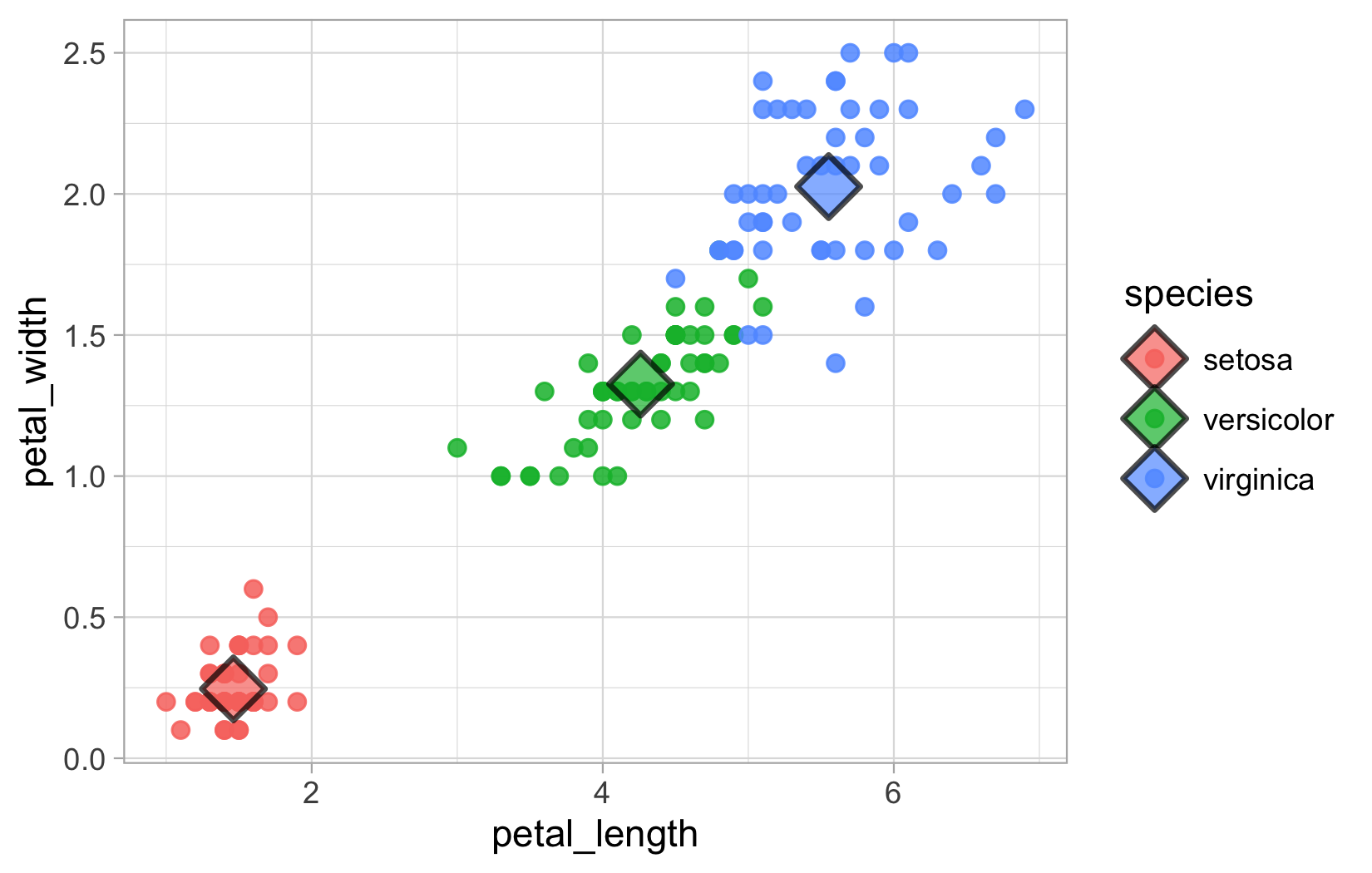


# MANOVA: Test Statistics

Several test statistics can be determined:

* Wilk’s λ: ratio of matrix determinants:|E|/|T|
* Smaller values: larger group differences
* Can be converted to approximate F ratios, compared to F distribution to find p

iris\_plot



|  |
| --- |
| MANOVA Assumptions |
| * Normal distribution:   + response vars should be normally distributed within groups (relatively robust - No outliers (use di2 to diagnose; very sensitive to this assumption   + Equal variance of the response variables across groups   + Linearity: response variables linearly related to each other   + No strong multicollinearity in response variables   + Best performance in balances designs |

# MANOVA: Model Fitting

iris\_manova\_model <- manova(cbind(sepal\_length, sepal\_width, petal\_length, petal\_width) ~ species, data = iris\_df)  
  
summary(iris\_manova\_model)

Df Pillai approx F num Df den Df Pr(>F)   
species 2 1.1919 53.466 8 290 < 2.2e-16 \*\*\*  
Residuals 147   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# MANOVA: Assumption Testing - Normality

Assumption test

1. Multivariate Normality

# Test multivariate normality for all data together  
response\_matrix <- iris\_df %>%   
 dplyr::select(sepal\_length, sepal\_width, petal\_length, petal\_width) %>%  
 as.matrix()  
  
# Multivariate Shapiro-Wilk test for entire dataset  
mshapiro.test(t(response\_matrix))

Shapiro-Wilk normality test  
  
data: Z  
W = 0.97935, p-value = 0.02342

# MANOVA: Assumption Testing - Homogeneity

Assumption test

1. Homogeneity of Covariance Matrices (Box’s M Test)

response\_vars <- iris\_df %>%   
 dplyr::select(sepal\_length, sepal\_width, petal\_length, petal\_width)   
  
# Box's M test for equality of covariance matrices  
box\_m\_result <- boxM(response\_vars, iris\_df$species)  
print(box\_m\_result)

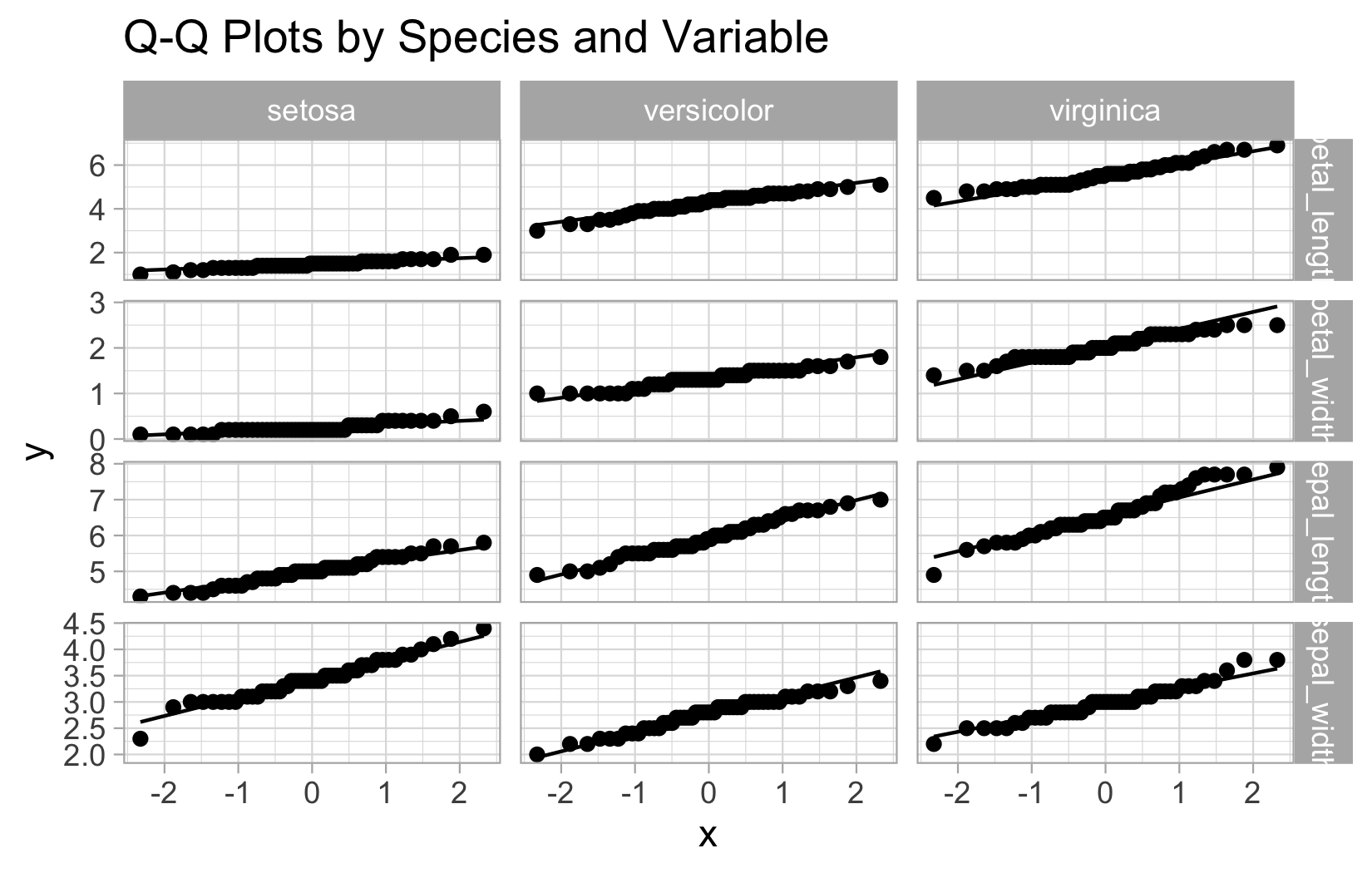
Box's M-test for Homogeneity of Covariance Matrices  
  
data: response\_vars  
Chi-Sq (approx.) = 140.94, df = 20, p-value < 2.2e-16

# MANOVA: Visual Assumption Assessment

Assumption test

1. Visual Assessment of Assumptions

# Create Q-Q plots for each variable by species  
iris\_long <- iris\_df %>%  
 pivot\_longer(cols = c(sepal\_length, sepal\_width, petal\_length, petal\_width),  
 names\_to = "variable",  
 values\_to = "value")  
  
iris\_long %>%  
 ggplot(aes(sample = value)) +  
 geom\_qq() +  
 geom\_qq\_line() +  
 facet\_grid(variable ~ species, scales = "free") +  
 labs(title = "Q-Q Plots by Species and Variable") +  
 theme\_light()



# MANOVA: Follow-up Univariate ANOVAs

## Follow-up Univariate ANOVAs

# Univariate ANOVAs for each response variable  
  
# Sepal Length ANOVA  
sepal\_length\_aov <- aov(sepal\_length ~ species, data = iris\_df)  
summary(sepal\_length\_aov)

Df Sum Sq Mean Sq F value Pr(>F)   
species 2 63.21 31.606 119.3 <2e-16 \*\*\*  
Residuals 147 38.96 0.265   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Sepal Width ANOVA  
sepal\_width\_aov <- aov(sepal\_width ~ species, data = iris\_df)  
summary(sepal\_width\_aov)

Df Sum Sq Mean Sq F value Pr(>F)   
species 2 11.35 5.672 49.16 <2e-16 \*\*\*  
Residuals 147 16.96 0.115   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Petal Length ANOVA  
petal\_length\_aov <- aov(petal\_length ~ species, data = iris\_df)  
summary(petal\_length\_aov)

Df Sum Sq Mean Sq F value Pr(>F)   
species 2 437.1 218.55 1180 <2e-16 \*\*\*  
Residuals 147 27.2 0.19   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Petal Width ANOVA  
petal\_width\_aov <- aov(petal\_width ~ species, data = iris\_df)  
summary(petal\_width\_aov)

Df Sum Sq Mean Sq F value Pr(>F)   
species 2 80.41 40.21 960 <2e-16 \*\*\*  
Residuals 147 6.16 0.04   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# MANOVA: Post-hoc Comparisons

## Post-hoc Comparisons using emmeans

# Sepal Length comparisons  
print("Sepal Length comparisons")

[1] "Sepal Length comparisons"

sepal\_length\_emm <- emmeans(sepal\_length\_aov, ~ species)  
pairs(sepal\_length\_emm)

contrast estimate SE df t.ratio p.value  
 setosa - versicolor -0.930 0.103 147 -9.033 <.0001  
 setosa - virginica -1.582 0.103 147 -15.366 <.0001  
 versicolor - virginica -0.652 0.103 147 -6.333 <.0001  
  
P value adjustment: tukey method for comparing a family of 3 estimates

# Sepal Width comparisons  
print("Sepal Width comparisons")

[1] "Sepal Width comparisons"

sepal\_width\_emm <- emmeans(sepal\_width\_aov, ~ species)  
pairs(sepal\_width\_emm)

contrast estimate SE df t.ratio p.value  
 setosa - versicolor 0.658 0.0679 147 9.685 <.0001  
 setosa - virginica 0.454 0.0679 147 6.683 <.0001  
 versicolor - virginica -0.204 0.0679 147 -3.003 0.0088  
  
P value adjustment: tukey method for comparing a family of 3 estimates

# Petal Length comparisons  
print("Petal Length comparisons")

[1] "Petal Length comparisons"

petal\_length\_emm <- emmeans(petal\_length\_aov, ~ species)  
pairs(petal\_length\_emm)

contrast estimate SE df t.ratio p.value  
 setosa - versicolor -2.80 0.0861 147 -32.510 <.0001  
 setosa - virginica -4.09 0.0861 147 -47.521 <.0001  
 versicolor - virginica -1.29 0.0861 147 -15.012 <.0001  
  
P value adjustment: tukey method for comparing a family of 3 estimates

# Petal Width comparisons  
print("Petal Width comparisons")

[1] "Petal Width comparisons"

petal\_width\_emm <- emmeans(petal\_width\_aov, ~ species)  
pairs(petal\_width\_emm)

contrast estimate SE df t.ratio p.value  
 setosa - versicolor -1.08 0.0409 147 -26.387 <.0001  
 setosa - virginica -1.78 0.0409 147 -43.489 <.0001  
 versicolor - virginica -0.70 0.0409 147 -17.102 <.0001  
  
P value adjustment: tukey method for comparing a family of 3 estimates

# Canonical Discriminant Analysis: Eigenvalues

## Eigenvalues and Canonical Variates

# Perform canonical discriminant analysis  
iris\_candisc <- candisc(iris\_manova\_model)  
  
# Display eigenvalues and canonical correlations  
cat("Canonical Discriminant Analysis Results:\n\n")

Canonical Discriminant Analysis Results:

print(iris\_candisc)

Canonical Discriminant Analysis for species:  
  
 CanRsq Eigenvalue Difference Percent Cumulative  
1 0.96987 32.19193 31.907 99.12126 99.121  
2 0.22203 0.28539 31.907 0.87874 100.000  
  
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 0.02344 199.145 8 288 < 2.2e-16 \*\*\*  
2 0.77797 13.794 3 145 5.794e-08 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Extract eigenvalues  
eigenvalues <- iris\_candisc$eigenvalues  
cat("\nEigenvalues:\n")

Eigenvalues:

print(eigenvalues)

[1] 3.219193e+01 2.853910e-01 -7.801056e-17 -1.398429e-15

# Calculate proportion of variance explained  
prop\_variance <- eigenvalues / sum(eigenvalues)  
cat("\nProportion of variance explained by each canonical variate:\n")

Proportion of variance explained by each canonical variate:

print(prop\_variance)

[1] 9.912126e-01 8.787395e-03 -2.402001e-18 -4.305862e-17

# Cumulative proportion  
cumulative\_prop <- cumsum(prop\_variance)  
cat("\nCumulative proportion of variance explained:\n")

Cumulative proportion of variance explained:

print(cumulative\_prop)

[1] 0.9912126 1.0000000 1.0000000 1.0000000

# Canonical Discriminant Analysis: Coefficients

## Canonical Coefficients (Eigenvectors)

# Display canonical coefficients (eigenvectors)  
cat("Raw Canonical Coefficients (Eigenvectors):\n")

Raw Canonical Coefficients (Eigenvectors):

print(iris\_candisc$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

cat("\nStandardized Canonical Coefficients:\n")

Standardized Canonical Coefficients:

print(iris\_candisc$coeffs.std)

Can1 Can2  
sepal\_length 0.4269548 0.01240753  
sepal\_width 0.5212417 0.73526131  
petal\_length -0.9472572 -0.40103782  
petal\_width -0.5751608 0.58103986

# Structure coefficients (correlations between original variables and canonical variates)  
cat("\nStructure Coefficients (Variable-Canonical Variate Correlations):\n")

Structure Coefficients (Variable-Canonical Variate Correlations):

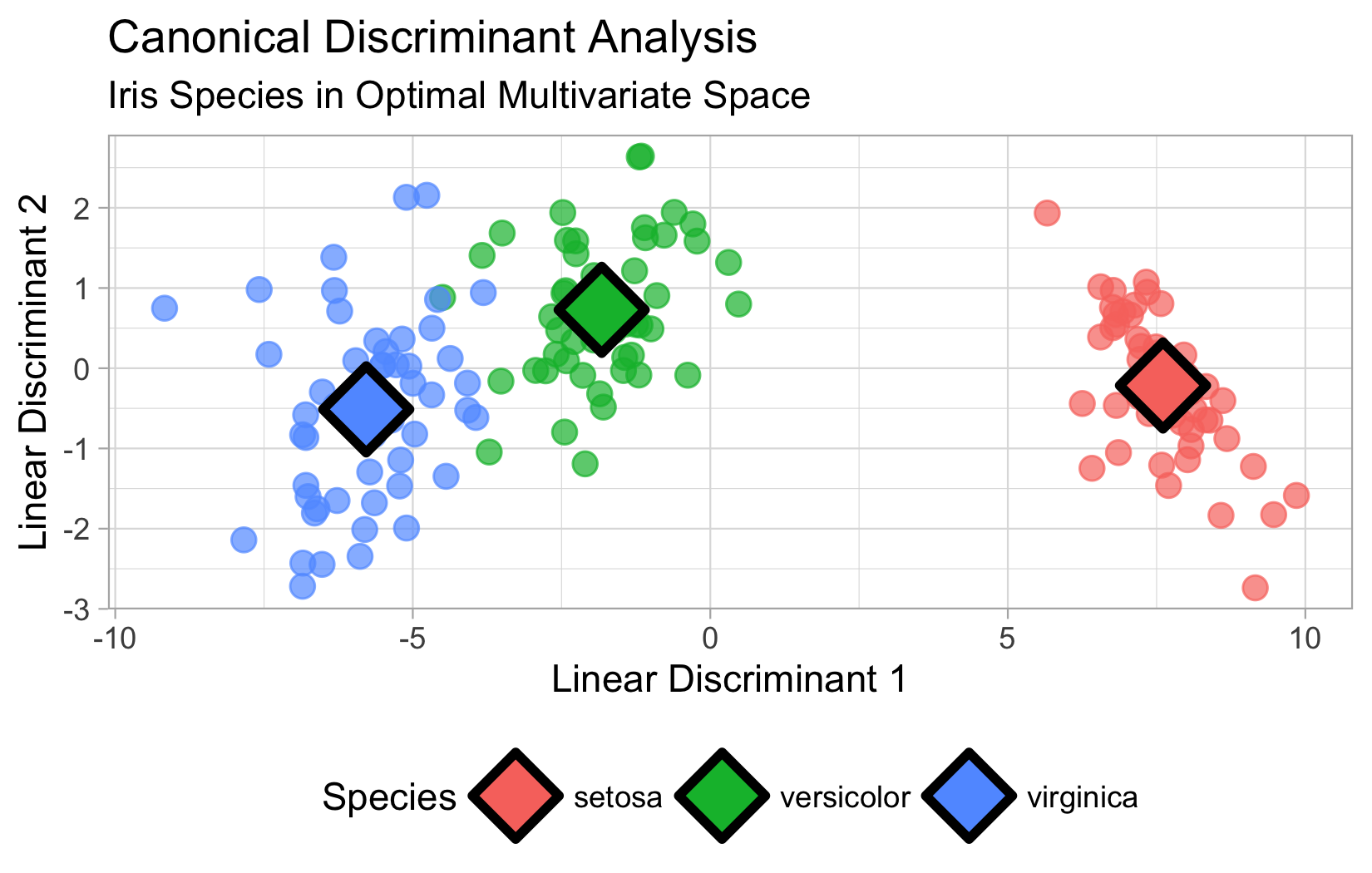
print(iris\_candisc$structure)

Can1 Can2  
sepal\_length -0.7918878 0.21759312  
sepal\_width 0.5307590 0.75798931  
petal\_length -0.9849513 0.04603709  
petal\_width -0.9728120 0.22290236

# Multivariate Visualization

Multivariate Visualization

# Extract canonical scores using the correct method  
  
# Alternative simpler approach - use lda from MASS package  
# library(MASS)  
iris\_lda <- MASS::lda(species ~ sepal\_length + sepal\_width + petal\_length + petal\_width, data = iris\_df)  
lda\_pred <- predict(iris\_lda)  
  
# Create dataframe with LDA scores (equivalent to canonical scores)  
canonical\_df\_plot <- data.frame(  
 Can1 = lda\_pred$x[, 1],  
 Can2 = lda\_pred$x[, 2],   
 species = iris\_df$species  
)  
  
# Calculate group centroids  
centroids\_plot <- canonical\_df\_plot %>%  
 group\_by(species) %>%  
 summarise(Can1\_mean = mean(Can1),  
 Can2\_mean = mean(Can2),  
 .groups = 'drop')  
  
# Create ggplot  
canonical\_df\_plot %>%  
 ggplot(aes(x = Can1, y = Can2, color = species)) +  
 geom\_point(size = 3, alpha = 0.7) +  
 geom\_point(data = centroids\_plot,  
 aes(x = Can1\_mean, y = Can2\_mean, fill = species),  
 shape = 23, color = "black", size = 8, stroke = 2) +  
 labs(title = "Canonical Discriminant Analysis",  
 subtitle = "Iris Species in Optimal Multivariate Space",  
 x = "Linear Discriminant 1",  
 y = "Linear Discriminant 2",  
 color = "Species",  
 fill = "Species") +  
 theme\_light() +  
 theme(legend.position = "bottom")



# MANOVA Results: Key Interpretation

### Interpretation of MANOVA

## Key Interpretation

**Pillai’s Trace (1.1919)**: This is large, indicating substantial group differences across the multivariate space.

**F-statistic (53.466)**: Very large F-value indicates strong evidence against the null hypothesis.

**P-value**: Essentially zero, meaning we reject the null hypothesis that all three species have the same multivariate means.

**Conclusion**: The three iris species are significantly different when considering all four morphological measurements simultaneously in multivariate space.

# MANOVA Test Results  
summary(iris\_manova\_model)

Df Pillai approx F num Df den Df Pr(>F)   
species 2 1.1919 53.466 8 290 < 2.2e-16 \*\*\*  
Residuals 147   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# MANOVA Results: Wilks’ Lambda

### Interpretation of MANOVA

**Wilks’ Lambda (0.023439)**: Very small value (close to 0) indicates:

* Only about 2.3% of the total variance is unexplained by group differences
* About 97.7% of the multivariate variance is explained by species differences
* Extremely strong group separation in multivariate space

**Effect Size**: Partial η² ≈ 1 - 0.023439 = 0.977 (very large effect size)

**F-statistic (199.15)**: Much larger than Pillai’s F-value because Wilks’ Lambda is often more powerful when assumptions are met

**Conclusion**: The three iris species show extremely large multivariate differences - they are very well separated in the 4-dimensional morphological space, with species explaining nearly 98% of the multivariate variance.

**Wilks’ vs Pillai’s**: Wilks’ Lambda is generally preferred when assumptions are met, while Pillai’s trace is more robust to assumption violations.

# Get Wilks' Lambda from manova for effect size  
manova\_summary <- summary(iris\_manova\_model, test = "Wilks")  
manova\_summary

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

# MANOVA Results: Effect Size

### Interpretation of MANOVA

**Meaning**: Approximately 97.7% of the total multivariate variance is explained by species differences.

**Effect Size Guidelines:** - **Small effect**: η² ≈ 0.01 (1% of variance explained) - **Medium effect**: η² ≈ 0.06 (6% of variance explained) - **Large effect**: η² ≈ 0.14 (14% of variance explained) - **Our result**: η² = 0.977 (**extremely large effect**)

**Practical Interpretation:** - Species are almost perfectly separated in multivariate morphological space - Only 2.3% of the variation in the four measurements is due to within-species differences - Species membership explains nearly all the multivariate variation - This represents one of the strongest group separations possible in real biological data

**Conclusion**: The iris species show dramatically different morphological profiles - they are essentially non-overlapping in the 4-dimensional space of sepal/petal measurements. This effect size indicates that species is an extremely powerful predictor of morphological characteristics.

# Effect Size (Partial Eta-squared approximation)  
wilks\_lambda <- manova\_summary$stats[1, "Wilks"]  
partial\_eta\_sq <- 1 - wilks\_lambda  
partial\_eta\_sq

[1] 0.9765614

# Canonical Variates: Variance Explained

### Interpretation of manova

**Element [1] = 0.9912**: - First canonical variate explains 99.12% of the between-group variance - This dimension captures almost all the multivariate group differences

**Element [2] = 0.0088**: - Second canonical variate explains 0.88% of the between-group variance - This dimension captures the remaining small group differences

## Interpretation

**Dimensionality**: The group differences are essentially **one-dimensional** - 99% of separation occurs along the first canonical axis.

**Biological Meaning**: There’s one primary “direction” in morphological space that best separates the three iris species, with a very minor secondary pattern.

**Practical Implication**: You could visualize almost all the group separation using just the first canonical variate, though plotting both dimensions shows the complete picture.

# Canonical Analysis Summary  
prop\_variance

[1] 9.912126e-01 8.787395e-03 -2.402001e-18 -4.305862e-17

sum(prop\_variance)

[1] 1

# Linear Discriminant Analysis: Detailed Results

### Interpretation of manova

Group means: - **Setosa**: Smallest overall, widest sepals, tiny petals - **Versicolor**: Medium-sized in most dimensions - **Virginica**: Largest overall, especially in petal dimensions - Clear size progression: setosa < versicolor < virginica

Coefficients of linear discriminants: - **LD1**: Positive weights for sepal measurements, negative for petal measurements - Separates small-petaled from large-petaled species - **LD2**: Mainly contrasts sepal width vs petal width - Fine-tunes separation between versicolor and virginica

Proportion of trace: - **LD1**: Explains 99.12% of between-group discrimination - **LD2**: Explains 0.88% of between-group discrimination - Confirms the separation is essentially one-dimensional (petal vs sepal contrast)

# LDA results for interpretation  
iris\_lda

Call:  
lda(species ~ sepal\_length + sepal\_width + petal\_length + petal\_width,   
 data = iris\_df)  
  
Prior probabilities of groups:  
 setosa versicolor virginica   
 0.3333333 0.3333333 0.3333333   
  
Group means:  
 sepal\_length sepal\_width petal\_length petal\_width  
setosa 5.006 3.428 1.462 0.246  
versicolor 5.936 2.770 4.260 1.326  
virginica 6.588 2.974 5.552 2.026  
  
Coefficients of linear discriminants:  
 LD1 LD2  
sepal\_length 0.8293776 -0.02410215  
sepal\_width 1.5344731 -2.16452123  
petal\_length -2.2012117 0.93192121  
petal\_width -2.8104603 -2.83918785  
  
Proportion of trace:  
 LD1 LD2   
0.9912 0.0088

# MANOVA Advantages over Multiple ANOVAs

## Advantages of MANOVA over Multiple ANOVAs

### Statistical Advantages

* Controls family-wise error rate (no need for Bonferroni correction)
* Accounts for correlations between response variables
* More powerful when variables are correlated
* Tests the ‘global’ null hypothesis

### Interpretational Advantages

* Reveals patterns in multivariate space that univariate tests miss
* Canonical variates show optimal linear combinations for group separation
* Provides insight into which variables work together to discriminate groups
* Shows the dimensionality of group differences

### Biological Relevance

* Organisms function as integrated wholes, not independent traits
* Natural selection acts on trait combinations, not isolated traits
* Multivariate approaches better reflect biological reality