FW 599 Special Topics: Multivariate Analysis of Ecological Data in R

Lecture 10: Statistical Inference – Part 1

Thursday, October 31, 2024

Lecture 10: Statistical Inference

- Principal Components Regression (PCR)
- Linear Discriminant Analysis (LDA)
- Permutation
- Multi-response Permutation Procedure (MRPP)

Recap: Indirect vs. Direct Comparison

Making Inferences from Ordination: Objectives

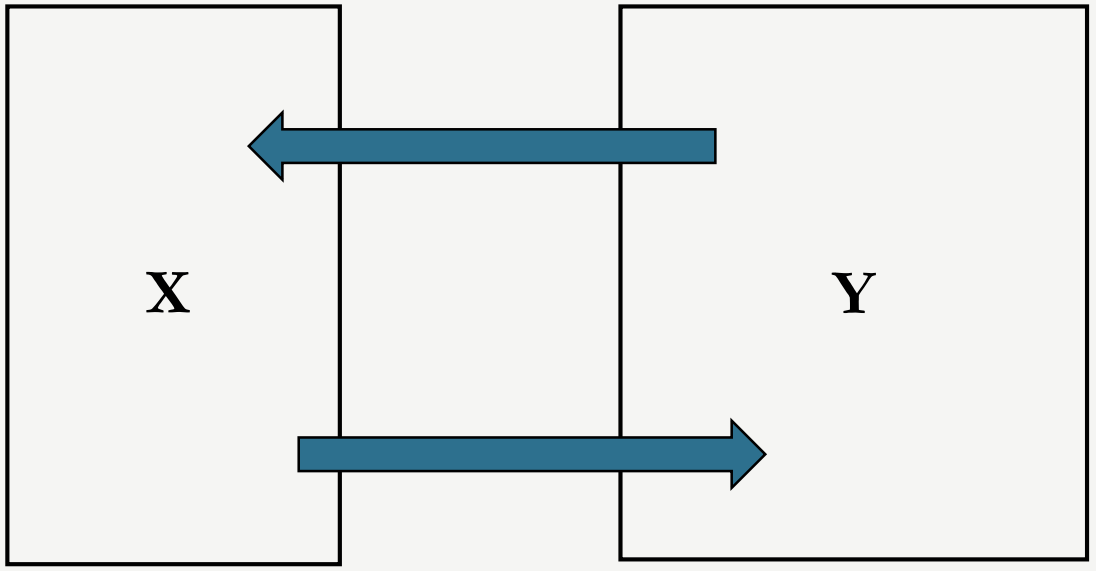
How do we translate our results into ecologically meaningful insights?

Interpretation: links patterns to ecological processes. Can be exploratory *or* inferential.

Inference: draws conclusions from patterns in complex datasets, usually to test hypotheses or identify key explanatory variables.

Making Inferences from Ordination: Direct Gradient Analysis

The goal of **direct comparison** is to simultaneously analyze the response and explanatory data matrices.



Making Inferences from Ordination: Explanatory

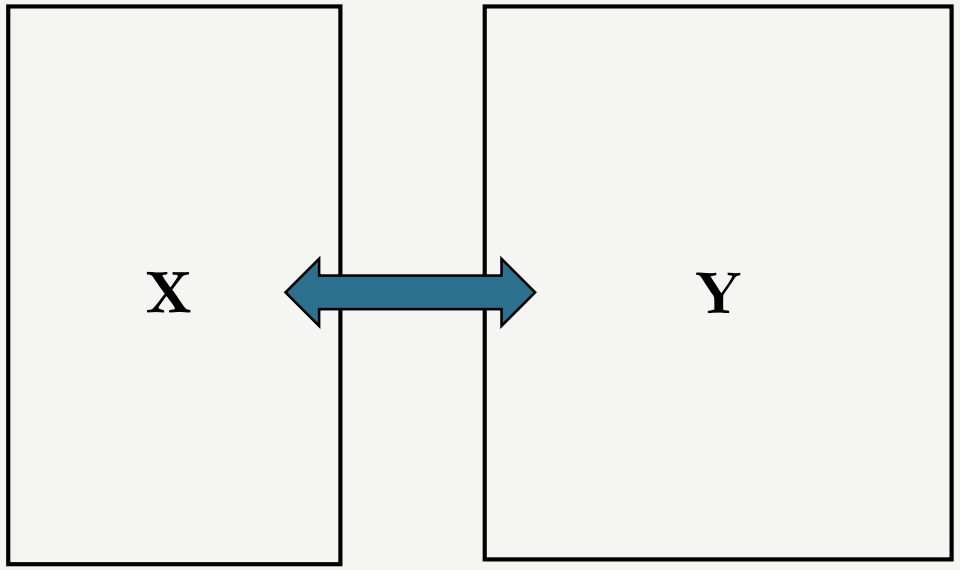
Explanatory data analysis looks for underlying relationships, patterns, and trends within a dataset.

1) **Indirect Comparison:** Treat principal axes/coordinates or clustering partitions as response variables in a regression analysis.

2) **Direct Comparison:** Redundancy Analysis (RDA) or Canonical Correspondence Analysis (CCA).

Canonical Methods: Constrained Ordination

Constrained ordination is an ordination technique in which the relationships between response variables and explanatory variables are explored.



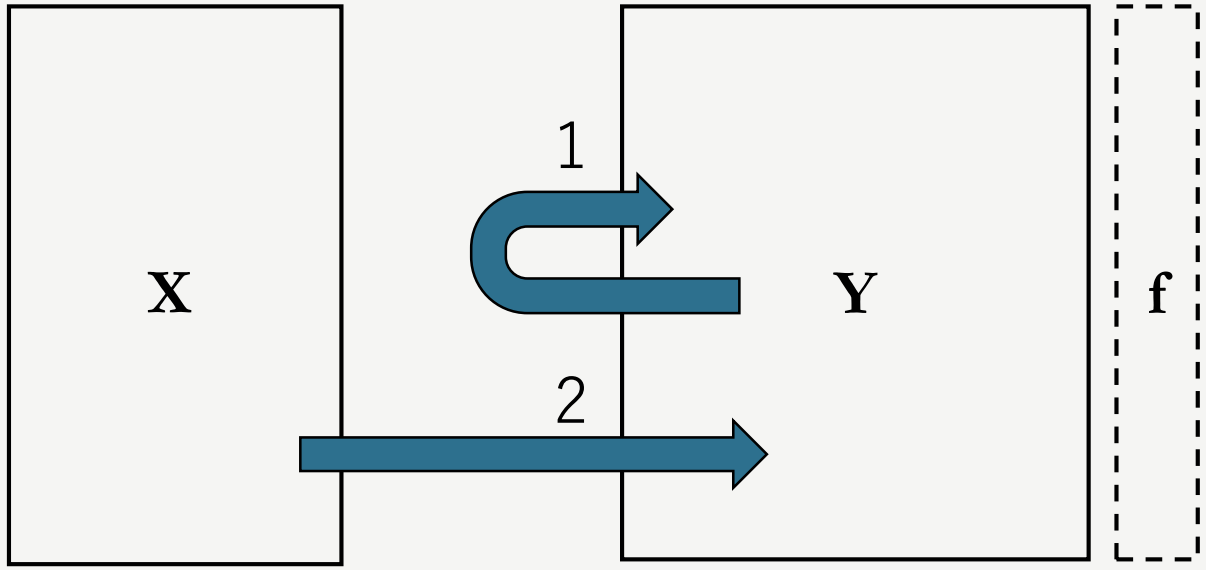
Canonical Methods: Constrained Ordination

Constrained ordination is an ordination technique in which the relationships between response variables and explanatory variables are explored.

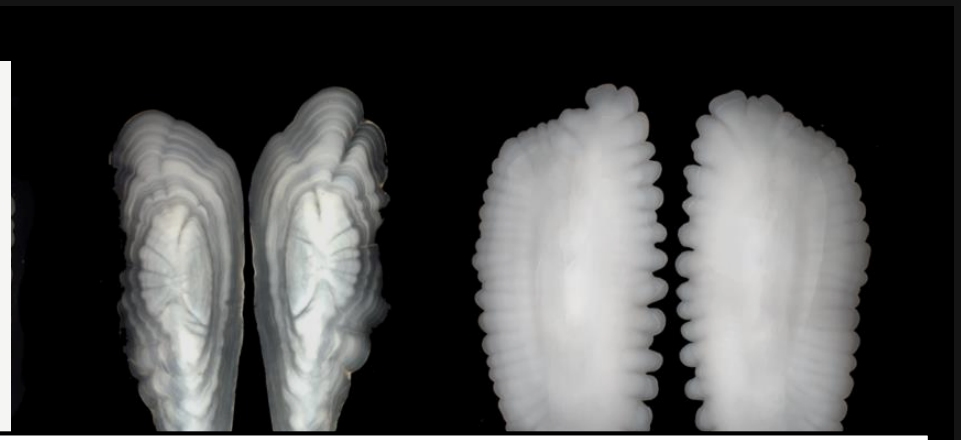
Explanatory variables **constrain** or guide the ordination by asking: how much of the variation in a multivariate dataset can be attributed to the explanatory variables?

Making Inferences from Ordination: Indirect Gradient Analysis

The goal of **indirect comparison** is to interpret the structure of the descriptors (response variables) using either the descriptors themselves or another set of descriptors.



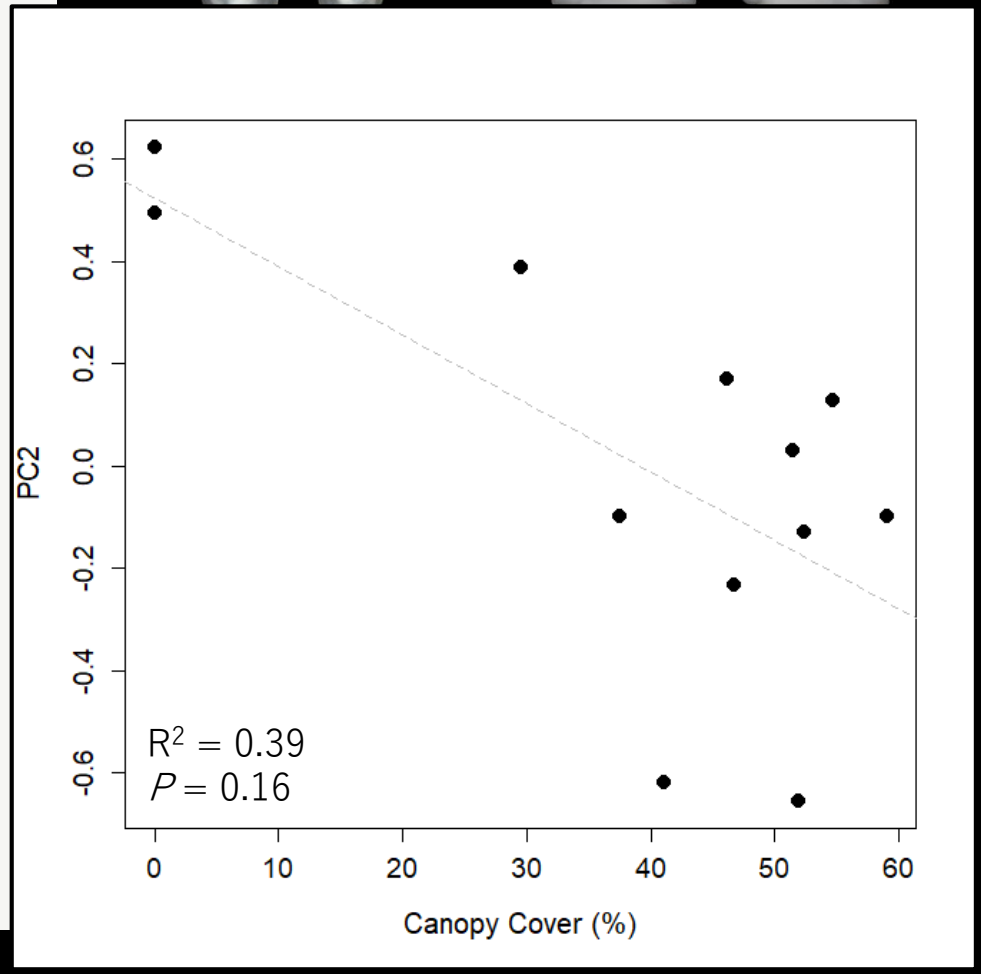
Making Inferences from Ordination: Explanatory



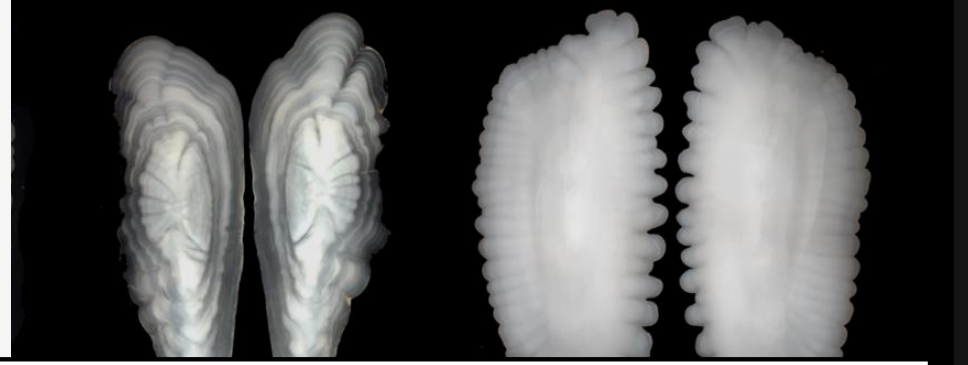
Explanatory data analysis looks for underlying relationships, patterns, and trends within a dataset.

1) **Indirect Comparison:** Treat principal axes/coordinates or clustering partitions as response variables in a regression analysis.

2) **Direct Comparison:** Redundancy Analysis (RDA) or Canonical Correspondence Analysis (CCA).



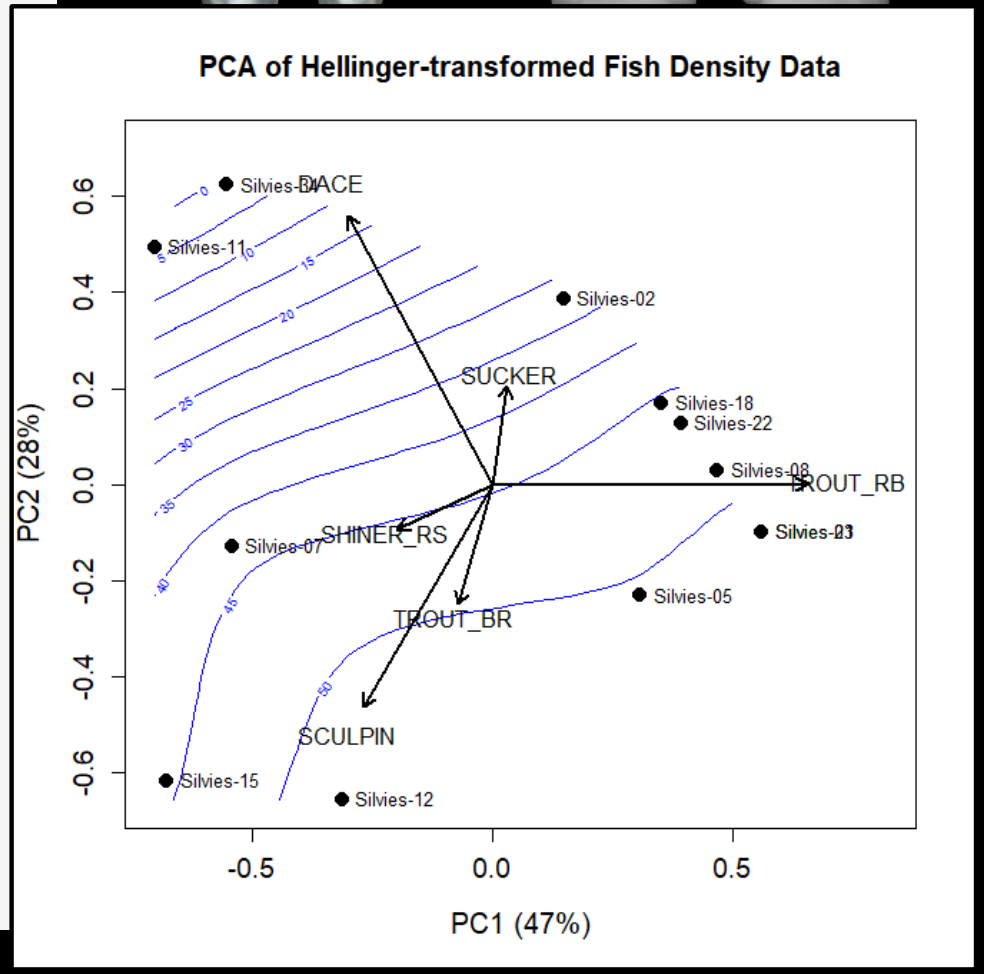
Making Inferences from Ordination: Explanatory



Explanatory data analysis looks for underlying relationships, patterns, and trends within a dataset.

1) **Indirect Comparison:** Treat principal axes/coordinates or clustering partitions as response variables in a regression analysis.

2) **Direct Comparison:** Redundancy Analysis (RDA) or Canonical Correspondence Analysis (CCA).



Indirect Comparison: Principal Components Regression

Indirect Comparison: Principal Components Regression

Principal Components Regression (**PCR**) is a regression technique that combines **Principal Component Analysis** (**PCA**) and **linear regression**.

Indirect Comparison: Principal Components Regression

Principal Components Regression (**PCR**) is a regression technique that combines **Principal Component Analysis** (**PCA**) and **linear regression**.

Goal is to reduce dimensionality of the data before fitting a linear regression model.

Indirect Comparison: Principal Components Regression

Principal Components Regression (**PCR**) is a regression technique that combines **Principal Component Analysis** (**PCA**) and **linear regression**.

- Solves issues with multicollinearity among predictor variables
- Improves model performance when working with a large number of correlated variables

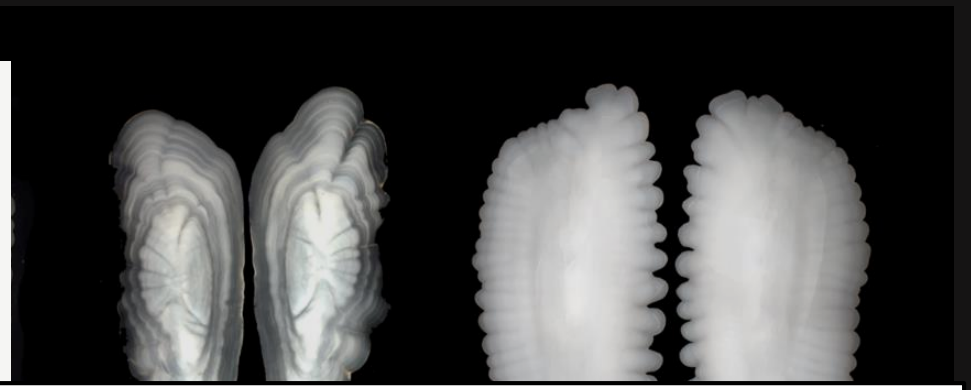
Indirect Comparison: Principal Components Regression

Step 1) Perform PCA

Indirect Comparison: Principal Components Regression

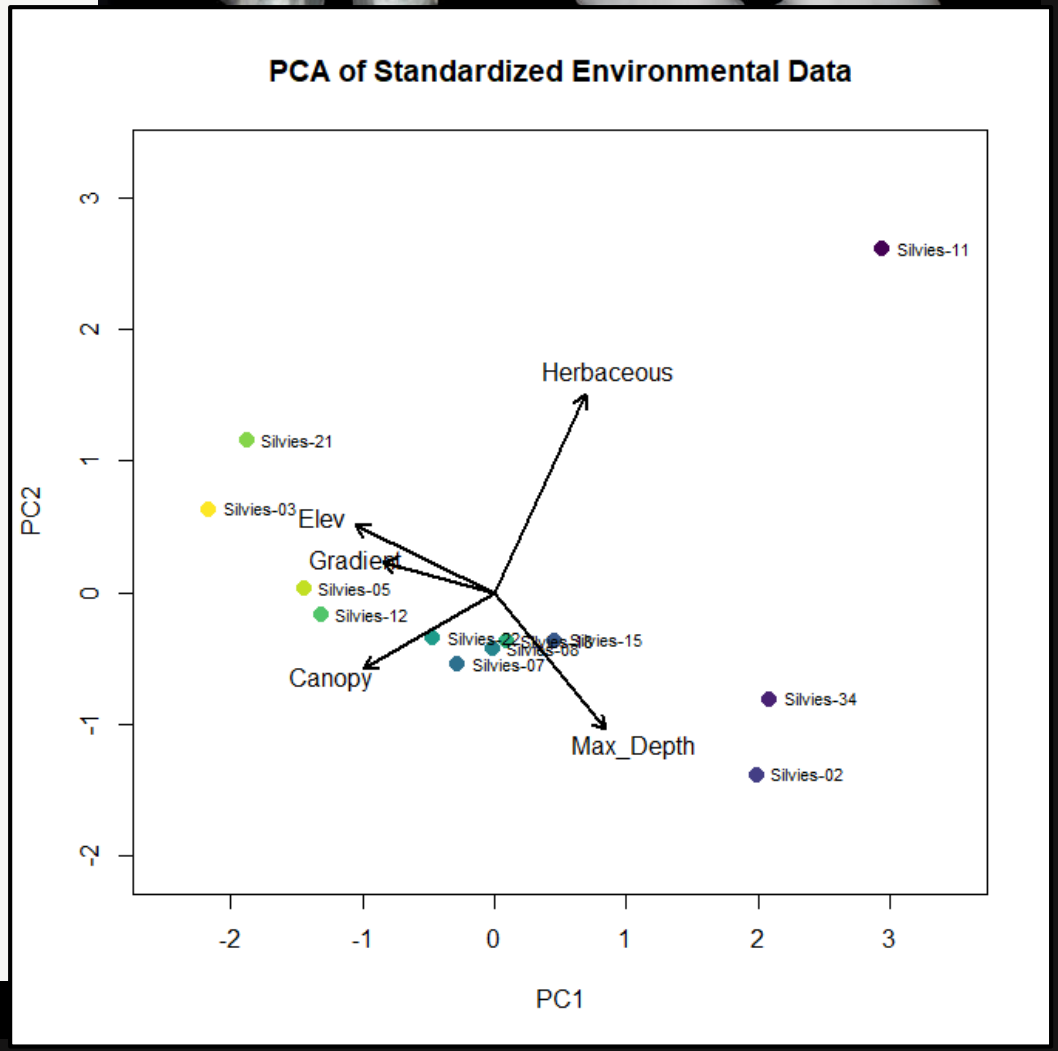
Site ID	Max Depth (m)	Gradient (%)	Elevation (m)	Canopy (%)	Herb (%)
Silvies-11	0.45	0.3	1439	0.0	55.1
Silvies-34	0.78	1.1	1487	0.0	0.0
Silvies-02	0.71	0.4	1372	29.6	0.0
Silvies-15	0.40	0.2	1471	41.1	0.0
Silvies-07	0.50	1.3	1547	52.3	0.0
Silvies-08	0.40	0.6	1492	51.4	0.0
Silvies-22	0.42	0.9	1555	54.7	0.0
Silvies-18	0.42	0.5	1510	46.2	0.0
Silvies-12	0.52	3.2	1658	51.9	0.0
Silvies-21	0.18	2.4	1713	37.5	0.0
Silvies-05	0.45	5.5	1565	46.7	0.0
Silvies-03	0.20	3.3	1634	59.0	0.0

Indirect Comparison: Principal Components Regression



Step 1) Perform PCA

- Decompose predictor variables into a set of uncorrelated principal components
- Each principal component is a linear (Euclidean) combination of the original variables

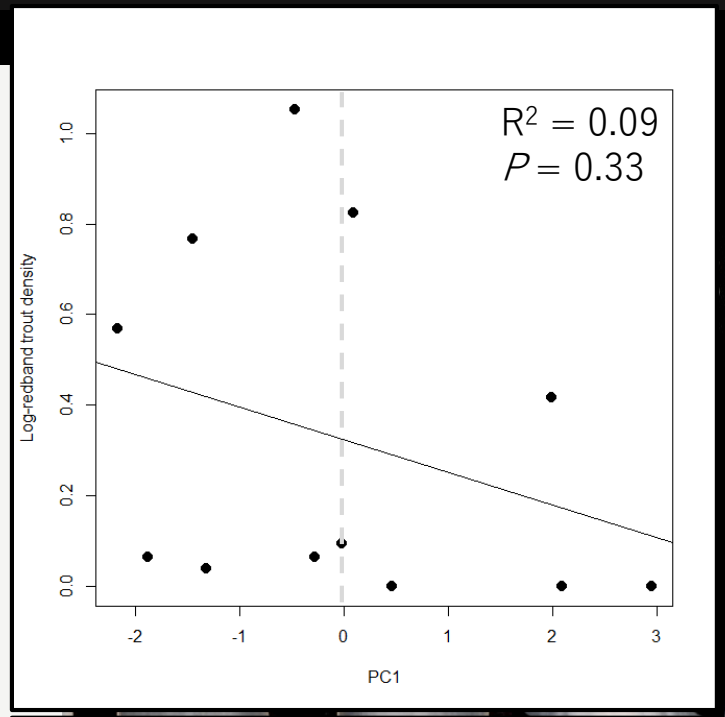


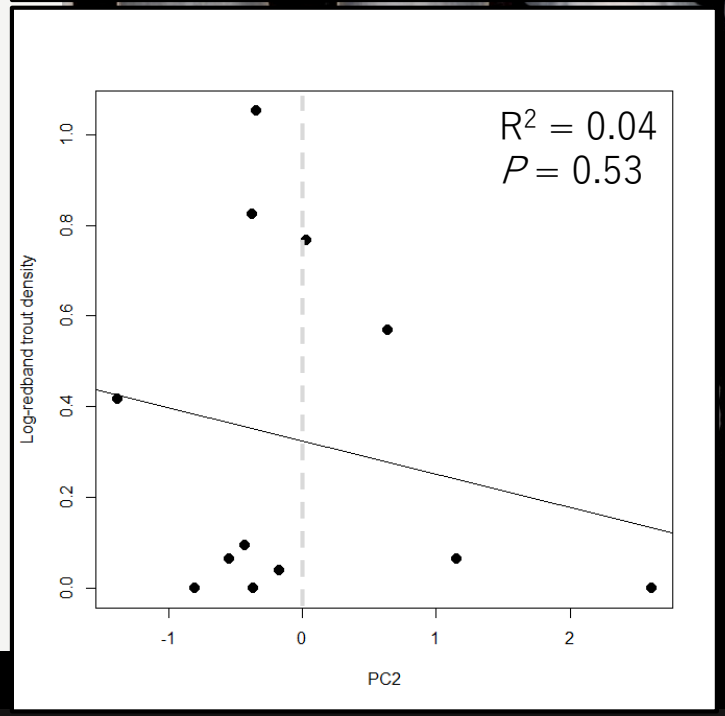
Indirect Comparison: Principal Components Regression

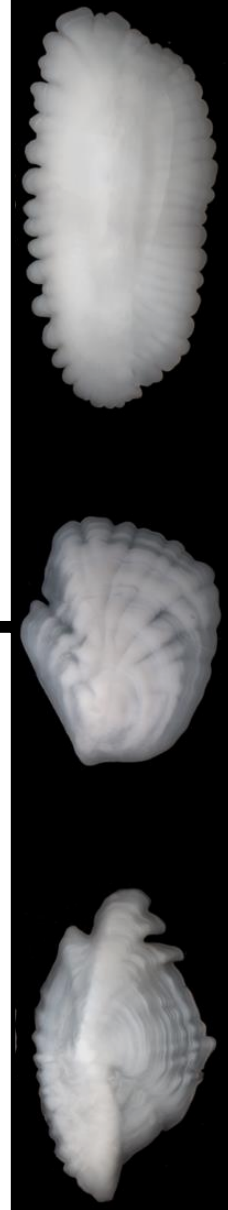
Step 2) Use the top principal components as predictors in a linear regression model

Indirect Comparison: Principal Components Regression

Step 2) Use the top principal components as predictors in a linear regression model

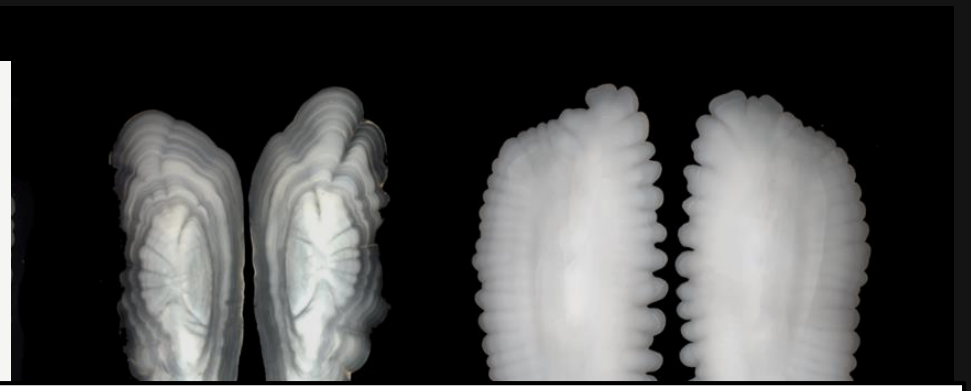


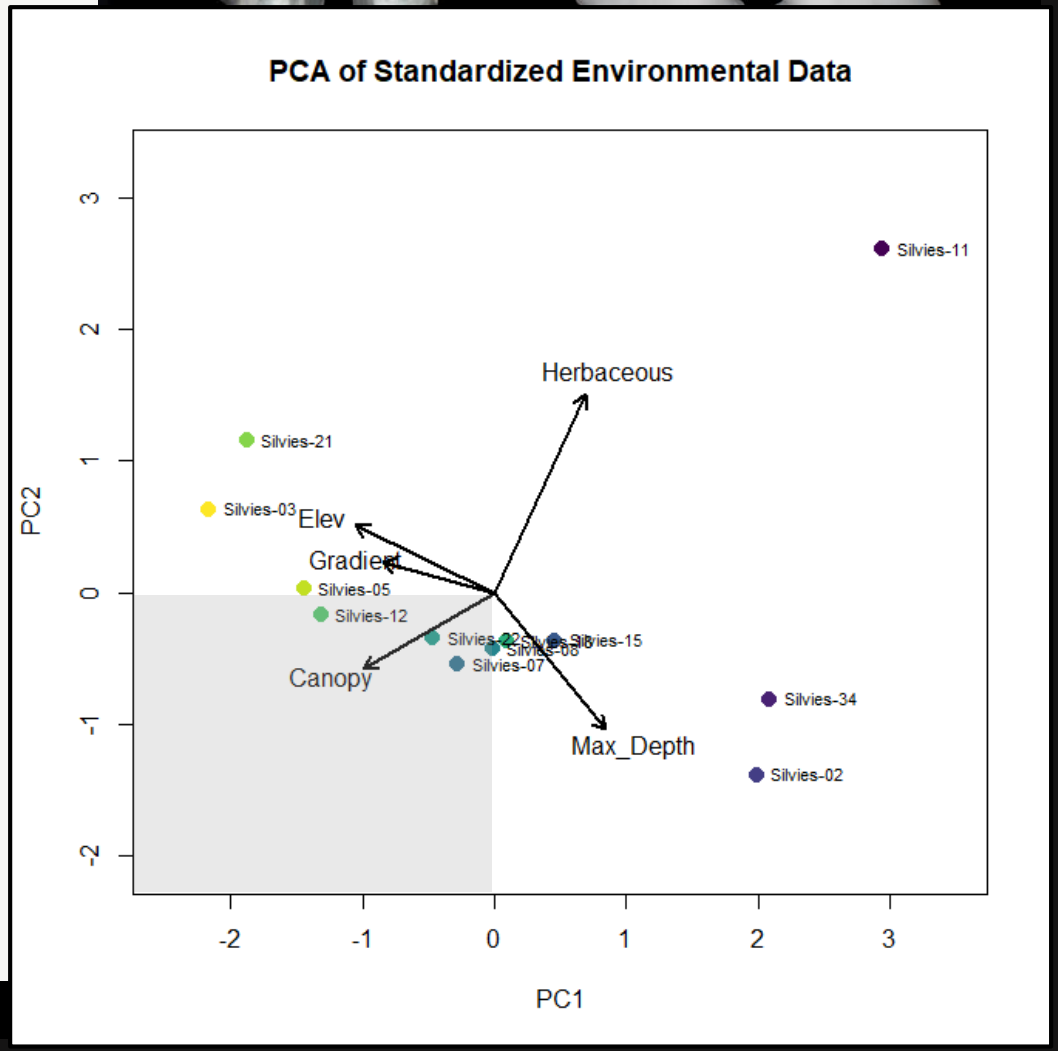




Indirect Comparison: Principal Components Regression

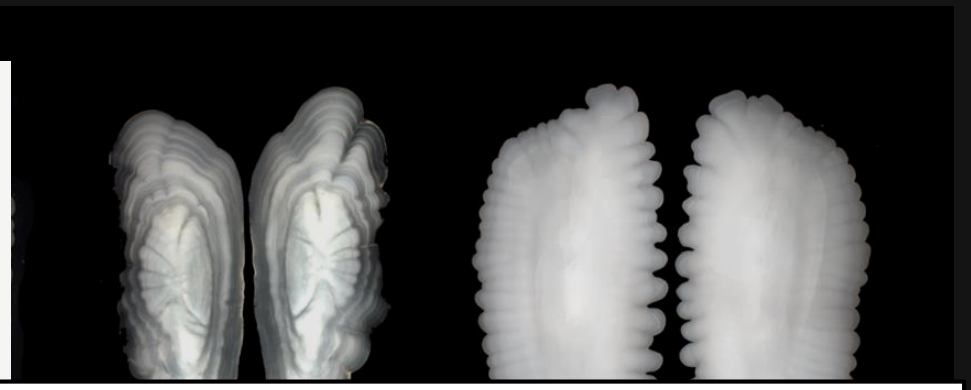
Step 2) Use the top principal components as predictors in a linear regression model

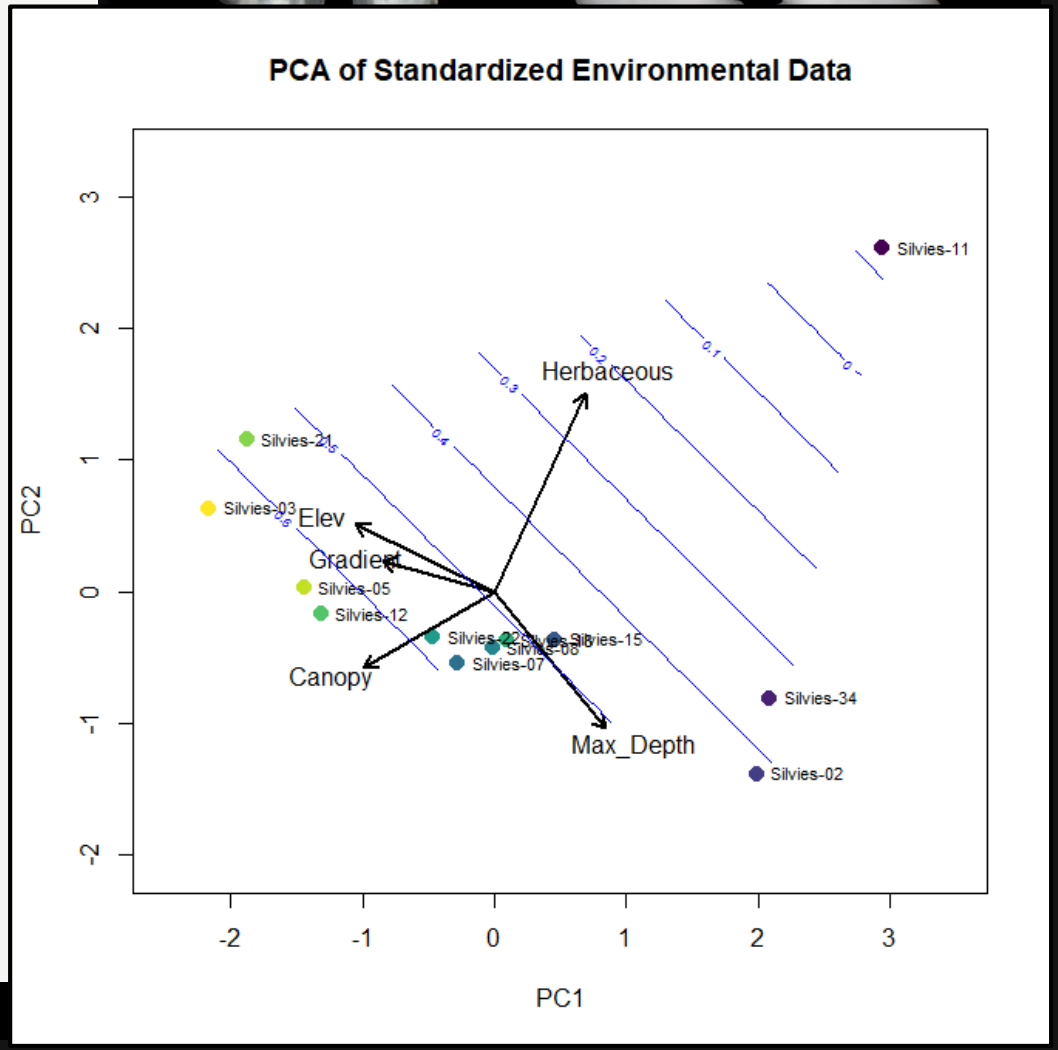




Indirect Comparison: Principal Components Regression

Step 2) Use the top principal components as predictors in a linear regression model





Indirect Comparison: Principal Components Regression

Advantages:

- Handles multicollinearity
- Reduces noise and redundancy
- Helps avoid overfitting

Indirect Comparison: Principal Components Regression

Advantages:

- Handles multicollinearity
- Reduces noise and redundancy
- Helps avoid overfitting

Disadvantages:

- Principal components may be challenging to interpret
- PCR doesn’t focus on predicting the dependent variable, but explaining the variance in the predictors

Indirect Comparison: Principal Components Regression

Advantages: Useful in a predictive capacity (e.g., species distribution modeling)

Indirect Comparison: Linear Discriminant Analysis

Indirect Comparison: Linear Discriminant Analysis

Discriminant Analysis or **Linear Discriminant Analysis (LDA)** is ordination technique that maximally separates a fixed (*a priori*) number of groups.

Indirect Comparison: Linear Discriminant Analysis

Discriminant Analysis or **Linear Discriminant Analysis (LDA)** is ordination technique that maximally separates a fixed (*a priori*) number of groups.

Goal is to find axes ("**discriminant functions**" or "**canonical axes**") that maximize among-group variation.

Indirect Comparison: Linear Discriminant Analysis

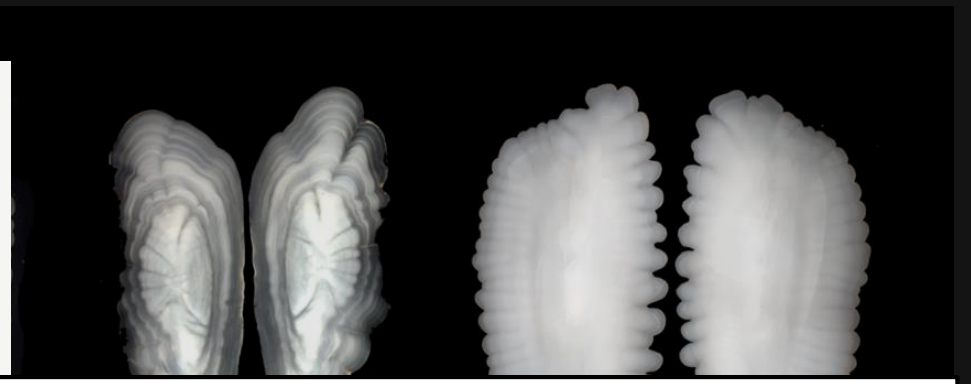
Discriminant Analysis or **Linear Discriminant Analysis (LDA)** is ordination technique that maximally separates a fixed (*a priori*) number of groups.

- Finds a linear combination of features that best separates two or more groups
- Projects data into reduced dimensional space

Indirect Comparison: Linear Discriminant Analysis

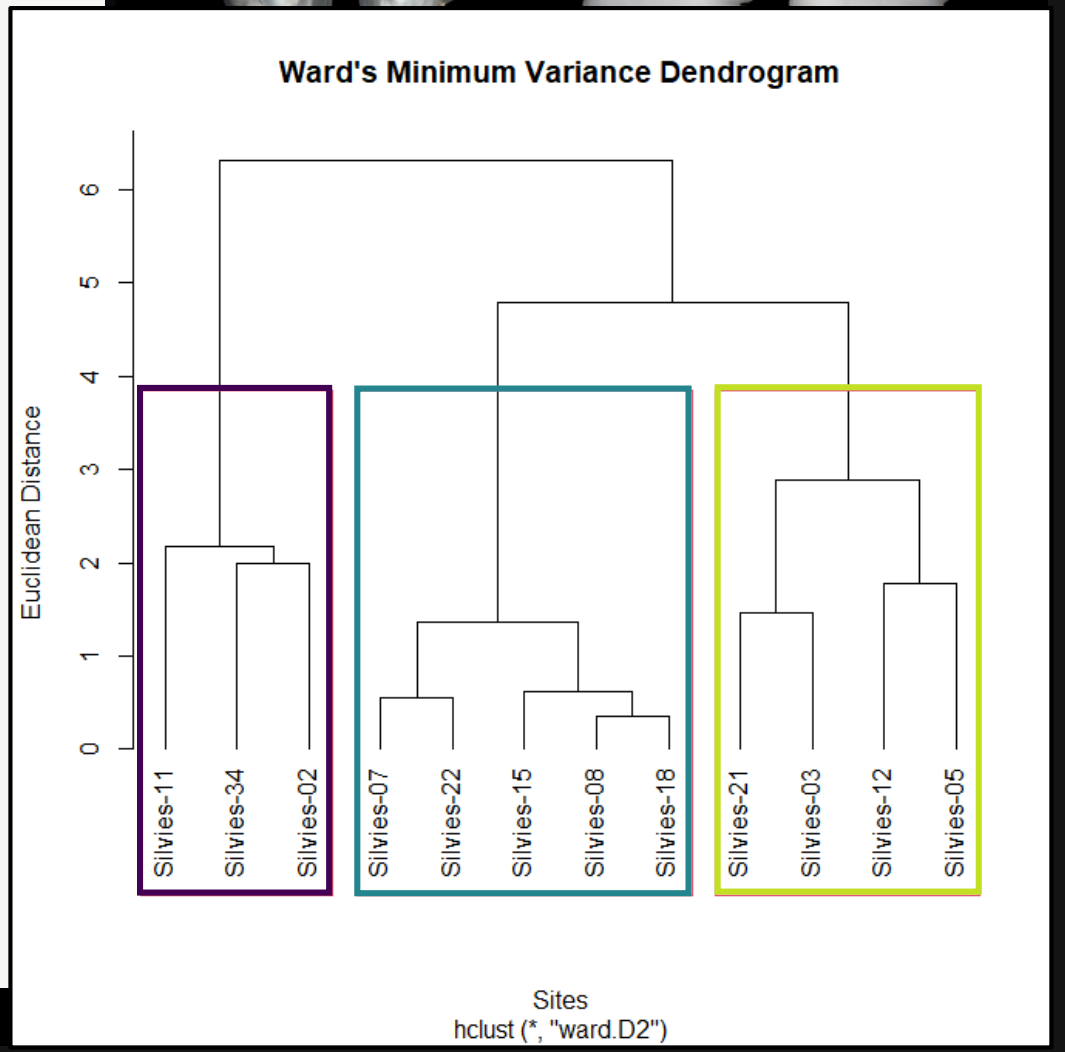
Step 1) Calculate the mean of each predictor variable for every group

Indirect Comparison: Linear Discriminant Analysis

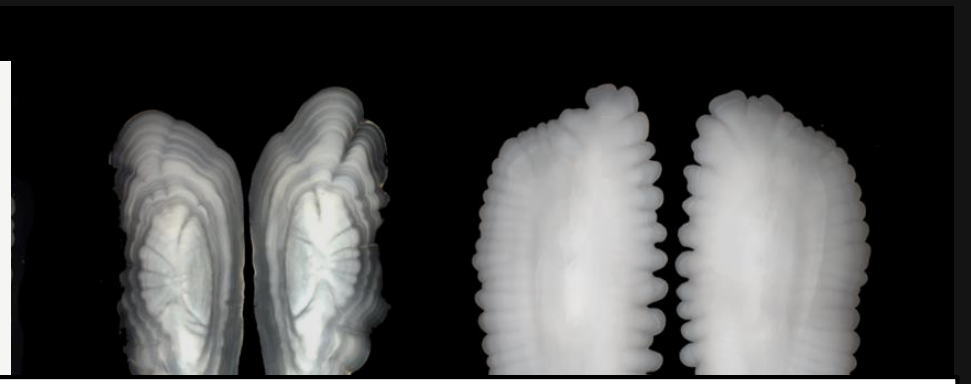


Step 1) Calculate the mean of each predictor variable for every group

Site ID	Max Depth (m)	Gradient (%)	Elevation (m)	Canopy (%)
GROUP 1	0.65	0.6	1433	9.9
GROUP 2	0.43	0.7	1515	49.1
GROUP 3	0.34	3.6	1642	48.8



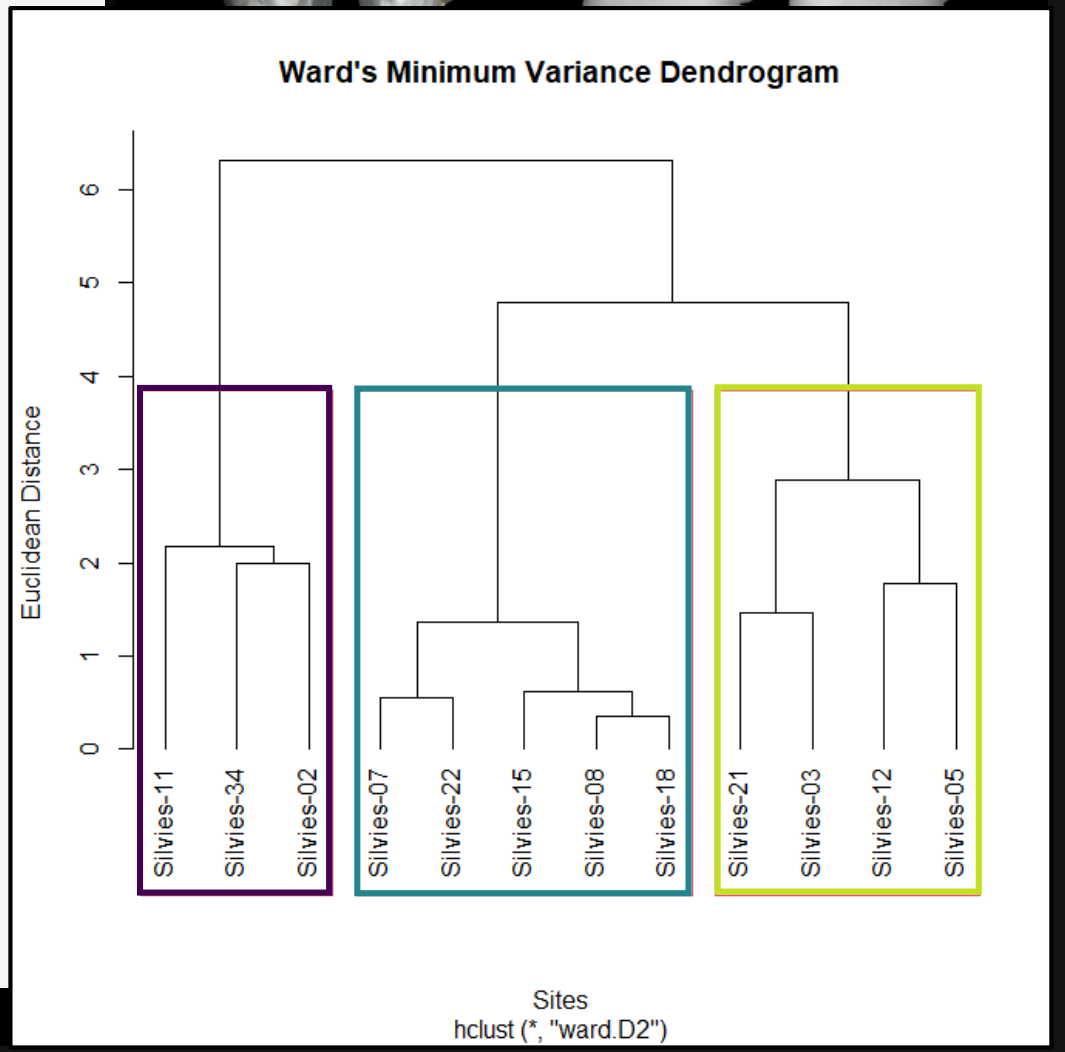
Indirect Comparison: Linear Discriminant Analysis



Step 1) Calculate the mean of each predictor variable for every group

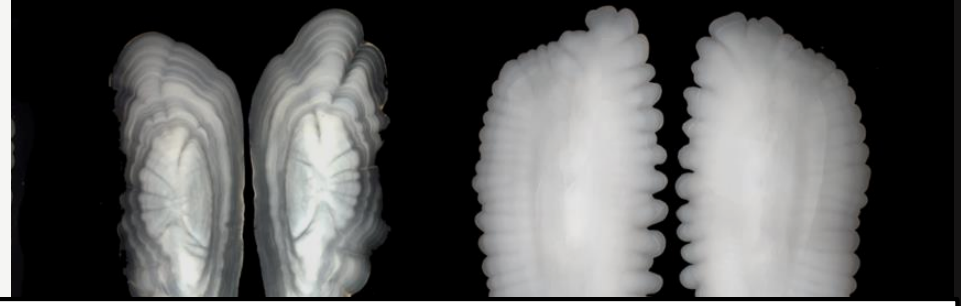
Site ID	Max Depth (m)	Gradient (%)	Elevation (m)	Canopy (%)
GROUP 1	0.65	0.6	1433	9.9
GROUP 2	0.43	0.7	1515	49.1
GROUP 3	0.34	3.6	1642	48.8

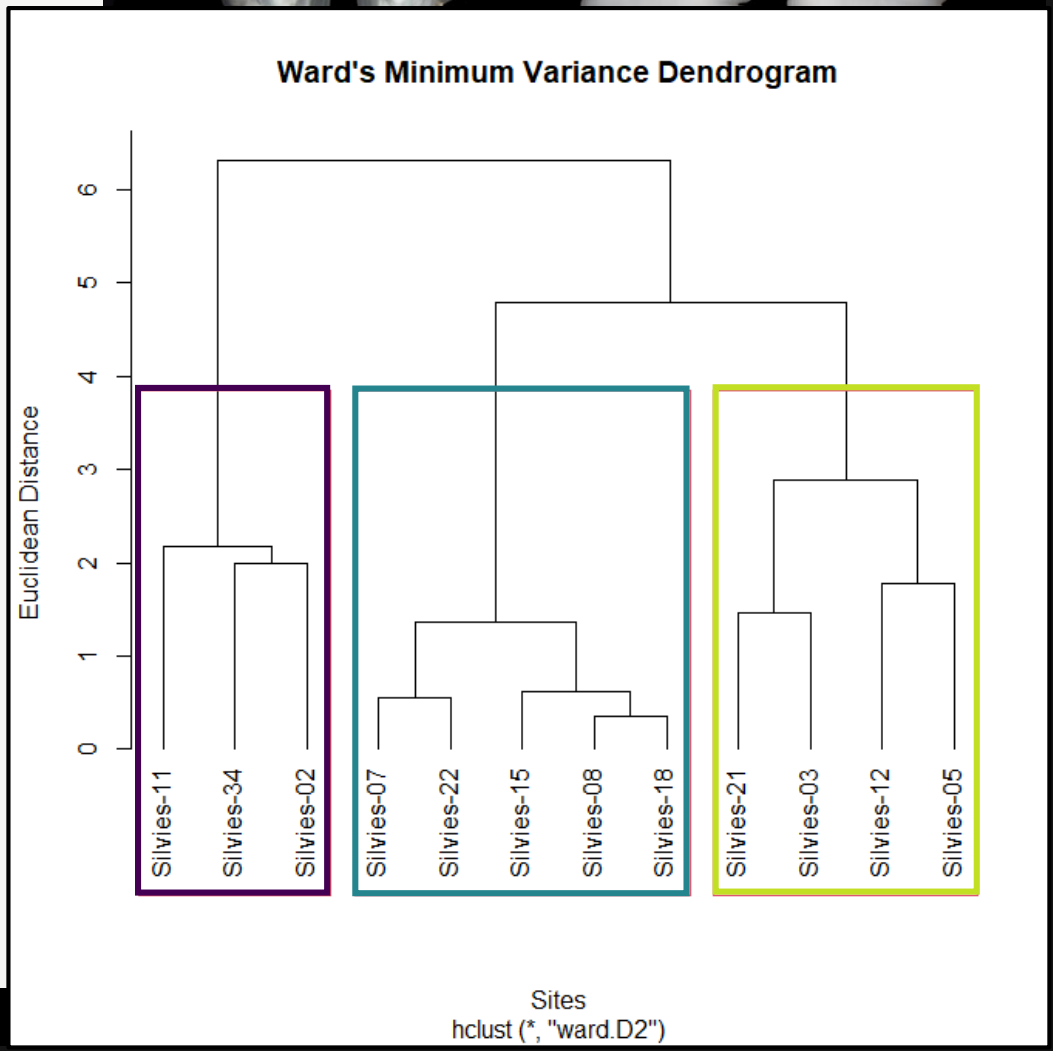
- Check for homogeneity of variance/group dispersions:
  - $P = 0.13$
- Distinct group means (Wilks test):
  - $P < 0.001$



Indirect Comparison: Linear Discriminant Analysis

Step 2) Calculate within ($\mathbf{S}_w$) and between ($\mathbf{S}_b$) group variances





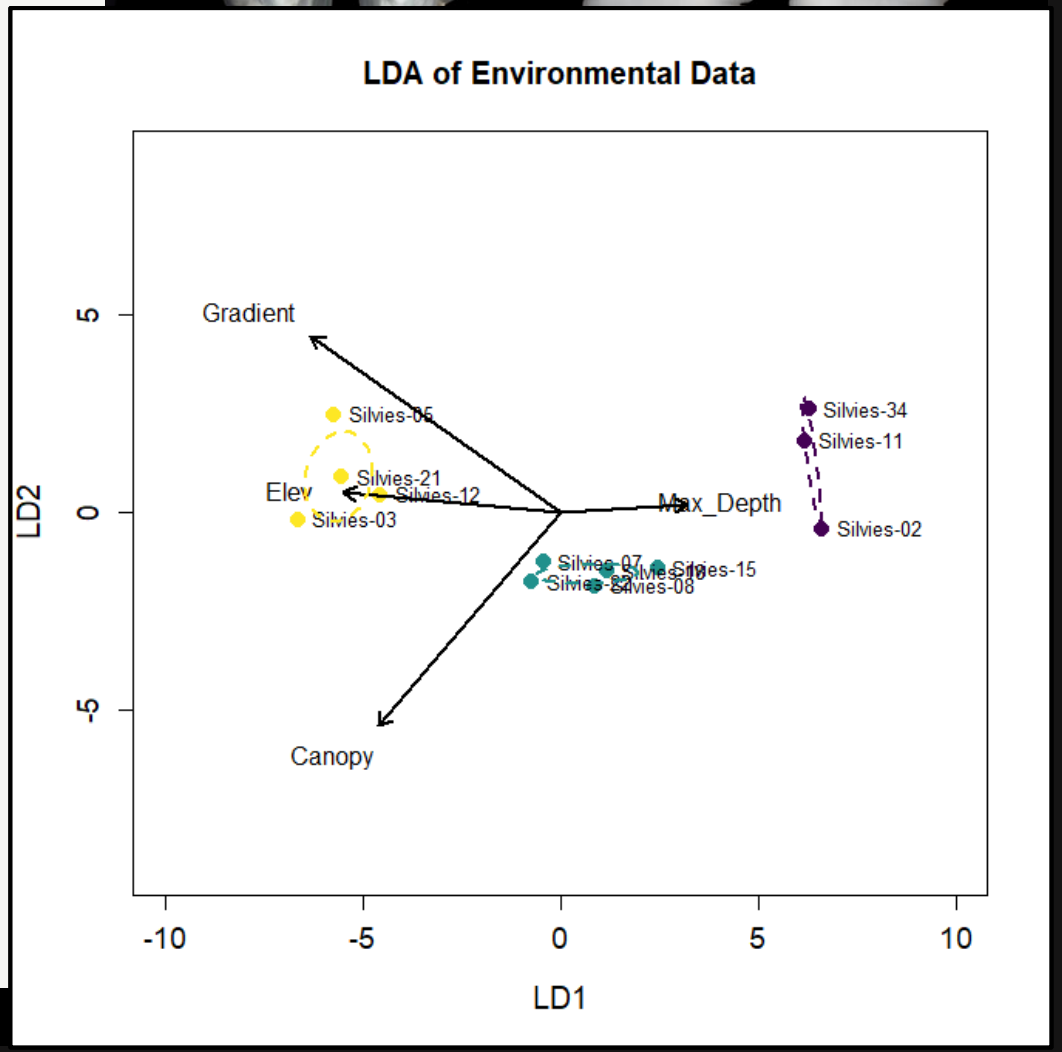
Indirect Comparison: Linear Discriminant Analysis

Step 2) Calculate within ($\mathbf{S}_w$) and between ($\mathbf{S}_b$) group variances

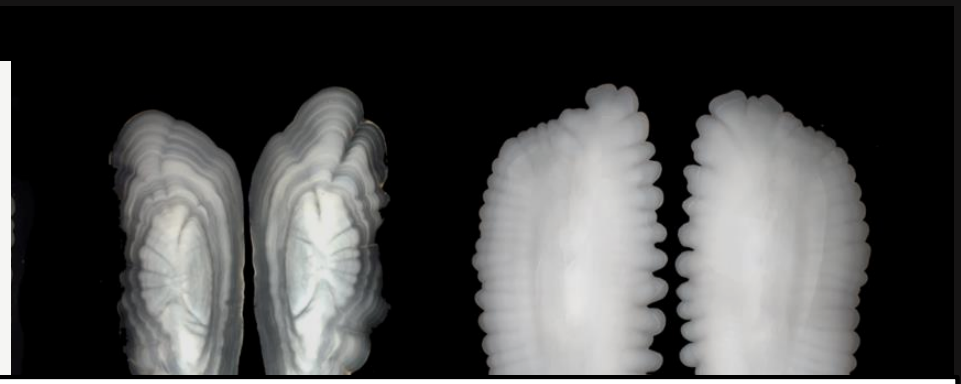
Step 3) Maximize class separation by solving the characteristic equation:

$$|\mathbf{S}_w^{-1}\mathbf{S}_b - \lambda \mathbf{I}| = 0$$

for eigenvectors, eigenvalues, and site scores

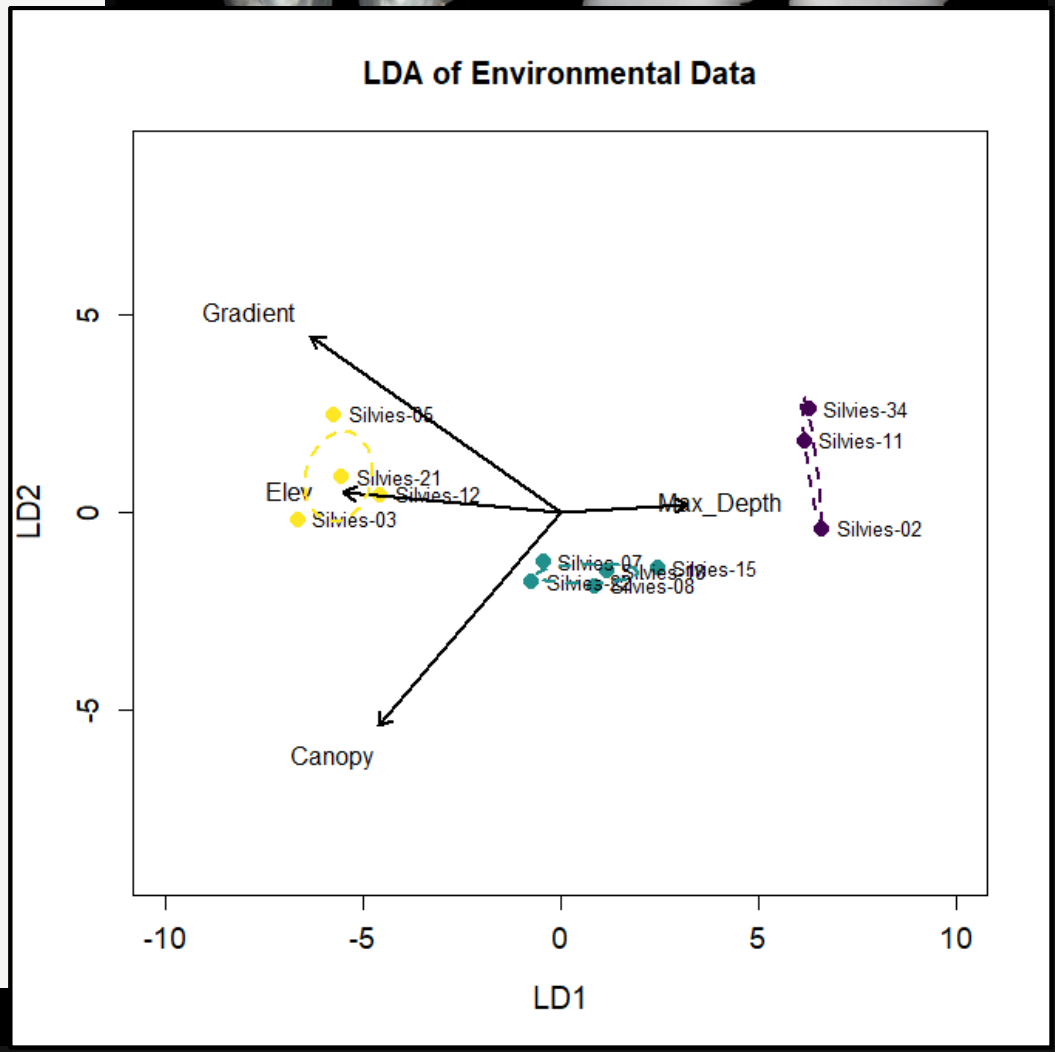


Indirect Comparison: Linear Discriminant Analysis



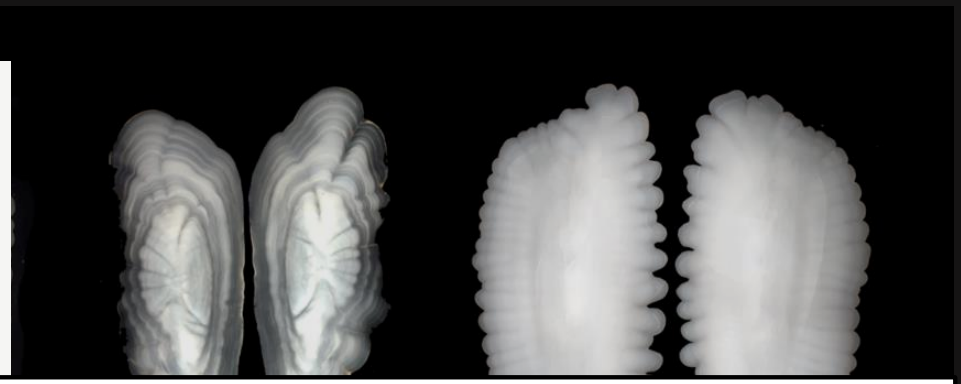
Check output using confusion matrix:

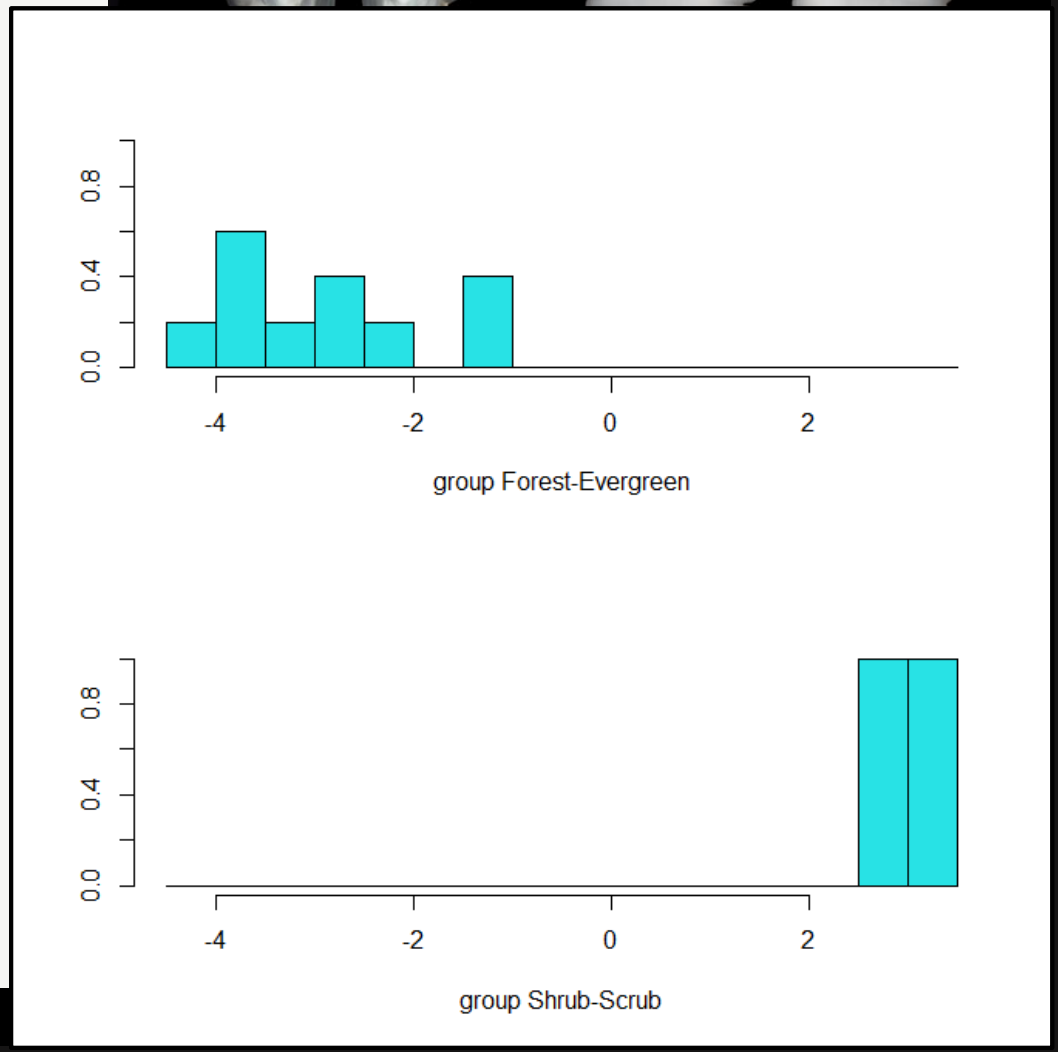
		True		
		1	**2**	**3**
Predicted	**1**	3	0	0
	2	0	5	0
	3	0	0	4



Indirect Comparison: Linear Discriminant Analysis

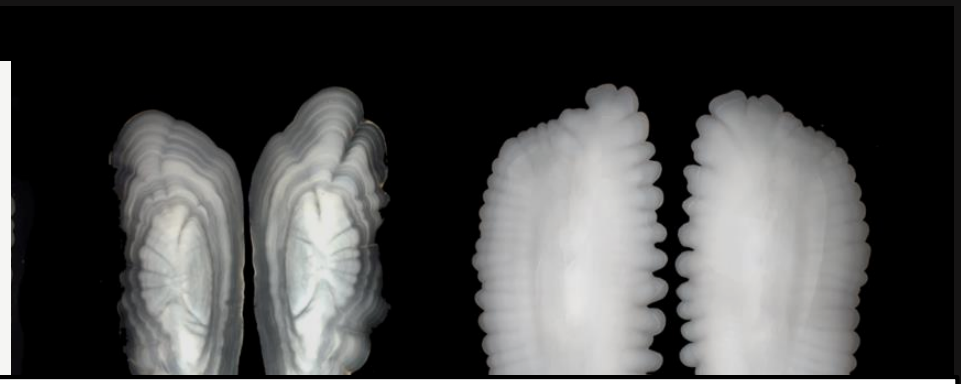
An LDA with only two classes/groups produces a single-dimensional output.

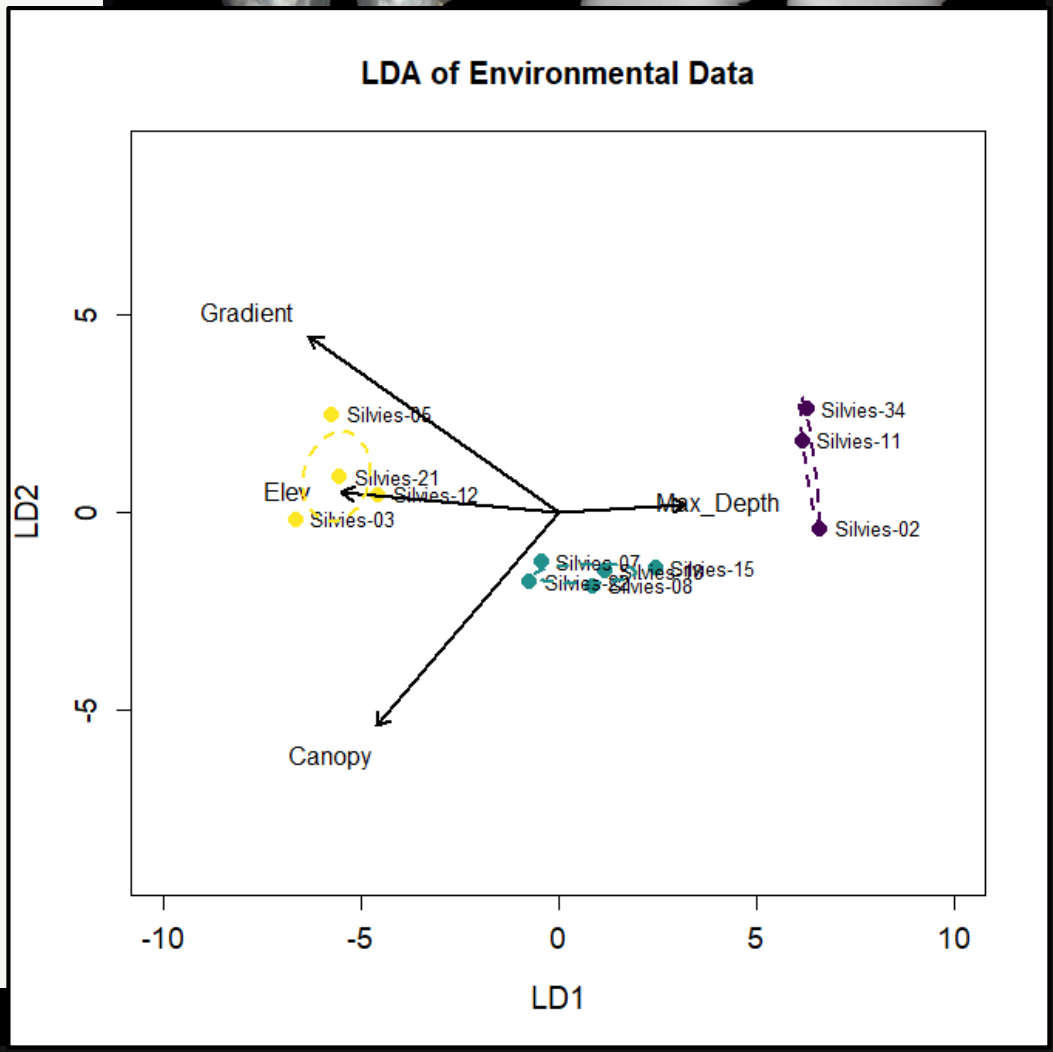




Indirect Comparison: Linear Discriminant Analysis

Step 4) Classify new objects (if that's the goal of your analysis)





Indirect Comparison: Linear Discriminant Analysis

Advantages:

- Efficient for multi-class classification problems
- Reduces dimensionality while preserving group-discriminatory information
- Performs well when distributions are close to normal

Indirect Comparison: Linear Discriminant Analysis

Advantages:

- Efficient for multi-class classification problems
- Reduces dimensionality while preserving group-discriminatory information
- Performs well when distributions are close to normal

Disadvantages:

- Assumptions similar to PCA, parametric methods
- May not capture complex, non-linear relationships among groups

Indirect Comparison: Permutation

Indirect Comparison: Permutation

A **permutation test** is a **non-parametric** method to assess the statistical significance of a test statistic by shuffling data and recalculating the statistic for each new arrangement.

Indirect Comparison: Permutation

A **permutation test** is a **non-parametric** method to assess the statistical significance of a test statistic by shuffling data and recalculating the statistic for each new arrangement.

- Instead of relying on assumptions of normality, uses the **distribution of the test statistic under the null hypothesis** generated by permuting the data.

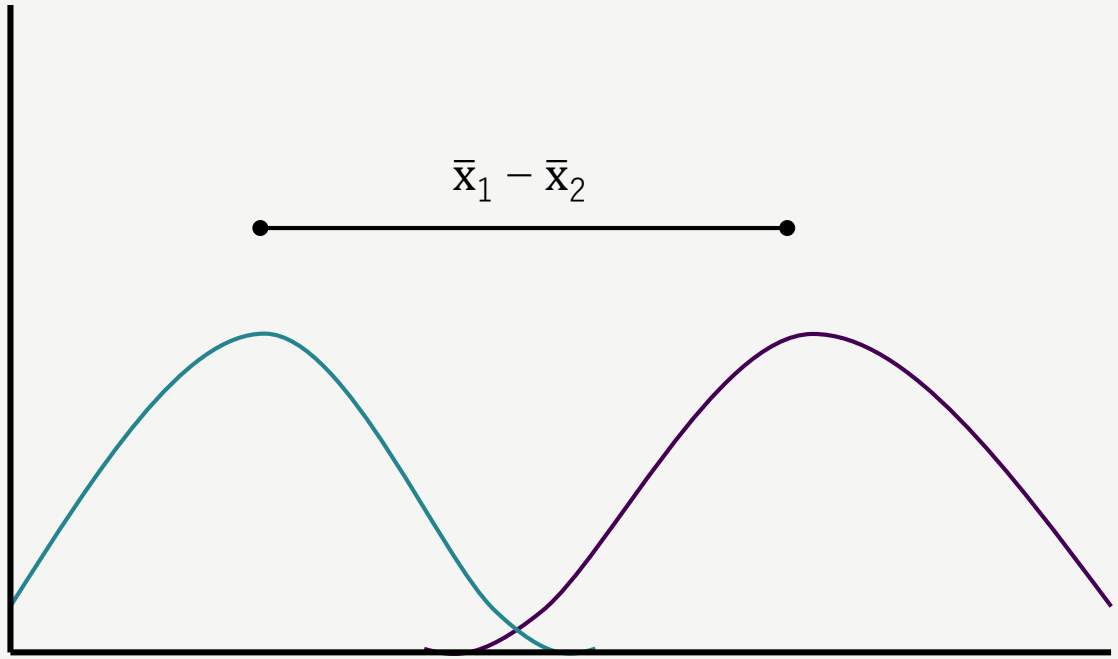
Indirect Comparison: Permutation

A **permutation test** is a **non-parametric** method to assess the statistical significance of a test statistic by shuffling data and recalculating the statistic for each new arrangement.

- Answers the question: “Is the observed result likely under random conditions?”

Indirect Comparison: Permutation

Step 1) Calculate the **test statistic** of interest for the original dataset (e.g., correlation coefficient, difference in group means)



Indirect Comparison: Permutation

Step 2) **Permute** or randomize the data

Indirect Comparison: Permutation

Step 2) **Permute** or randomize the data

- Randomly shuffle a vector of continuous variables
- Randomly assign values to classes/groups

Indirect Comparison: Permutation

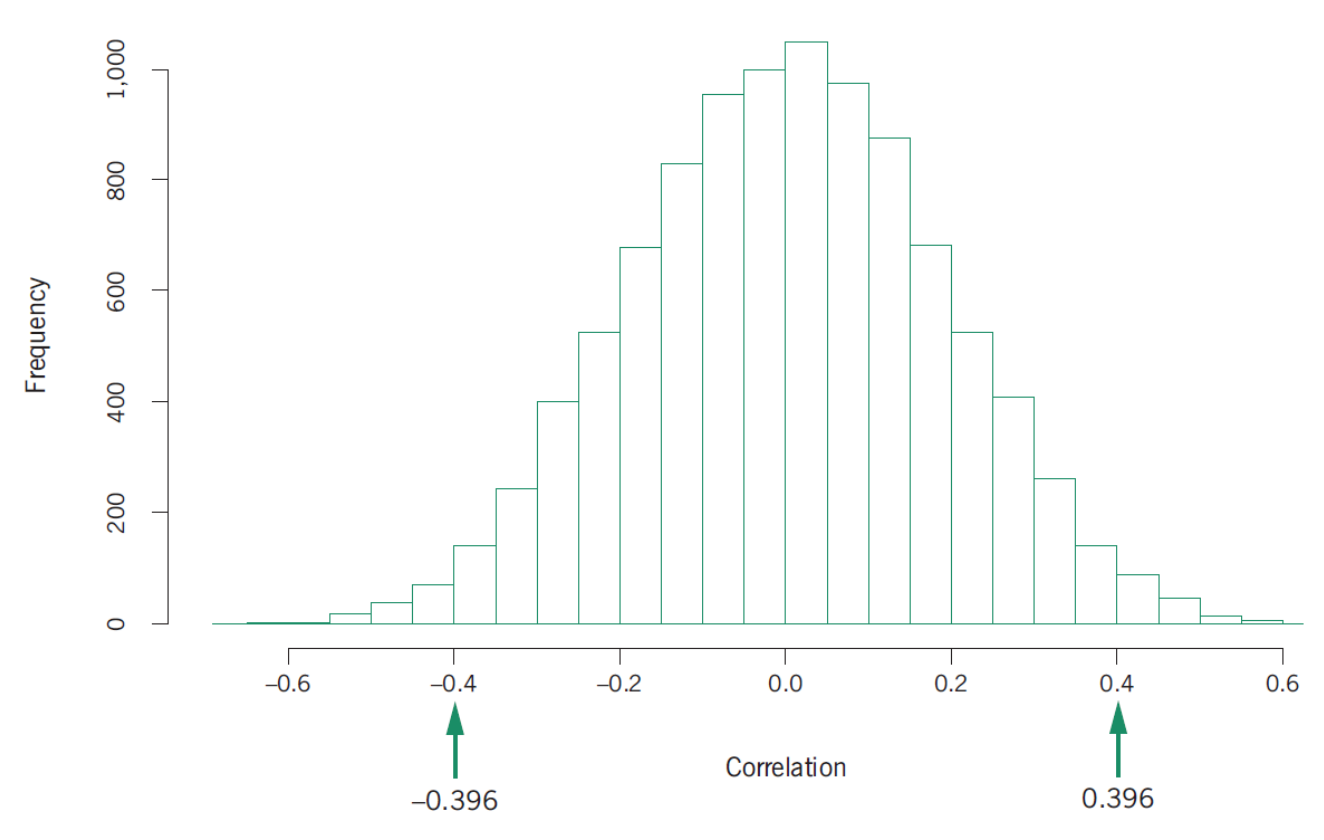
Step 2) **Permute** or randomize the data

- Randomly shuffle a vector of continuous variables
- Randomly assign values to classes/groups

Step 3) For **each** permutation, recalculate the test statistic

Indirect Comparison: Permutation

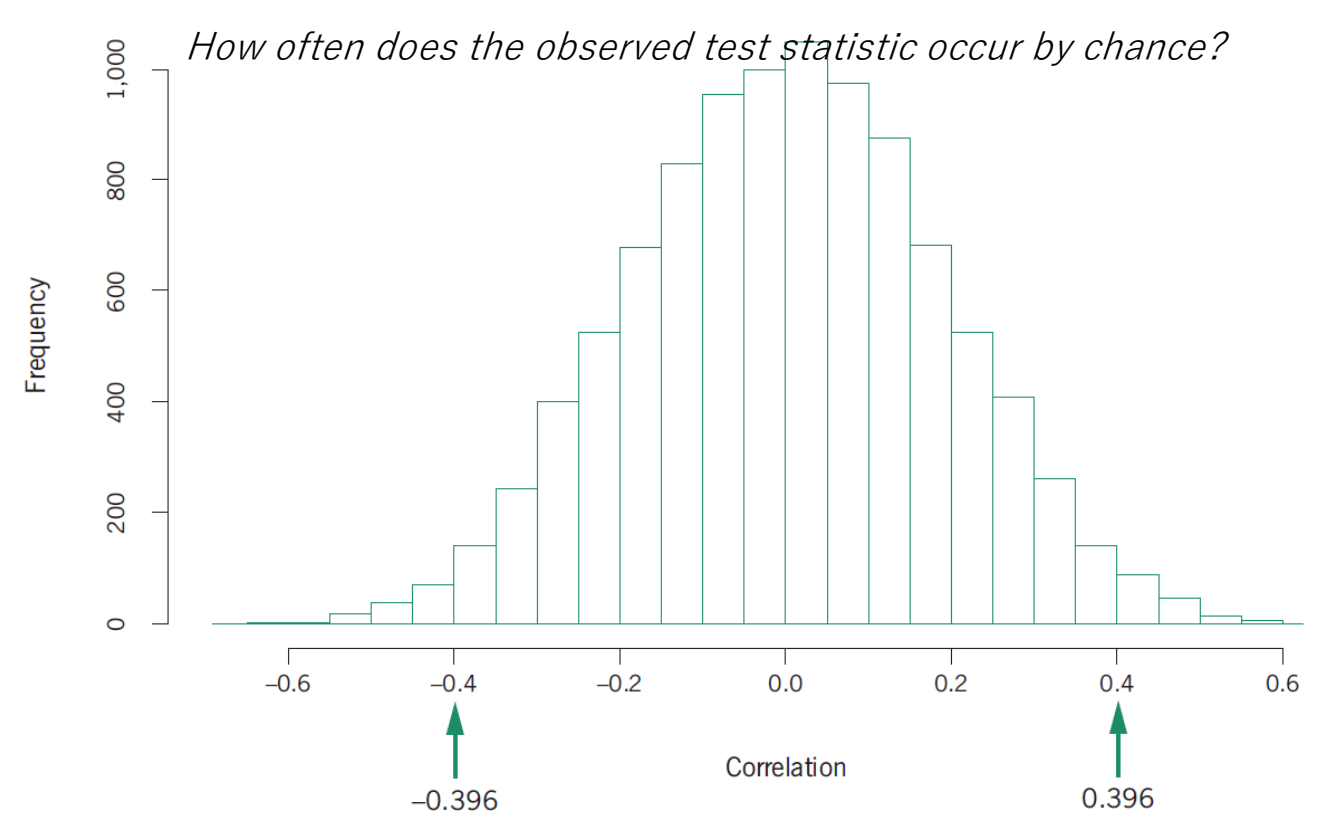
Step 4) Generate the permutation distribution



Greenacre & Primicerio 17.3

Indirect Comparison: Permutation

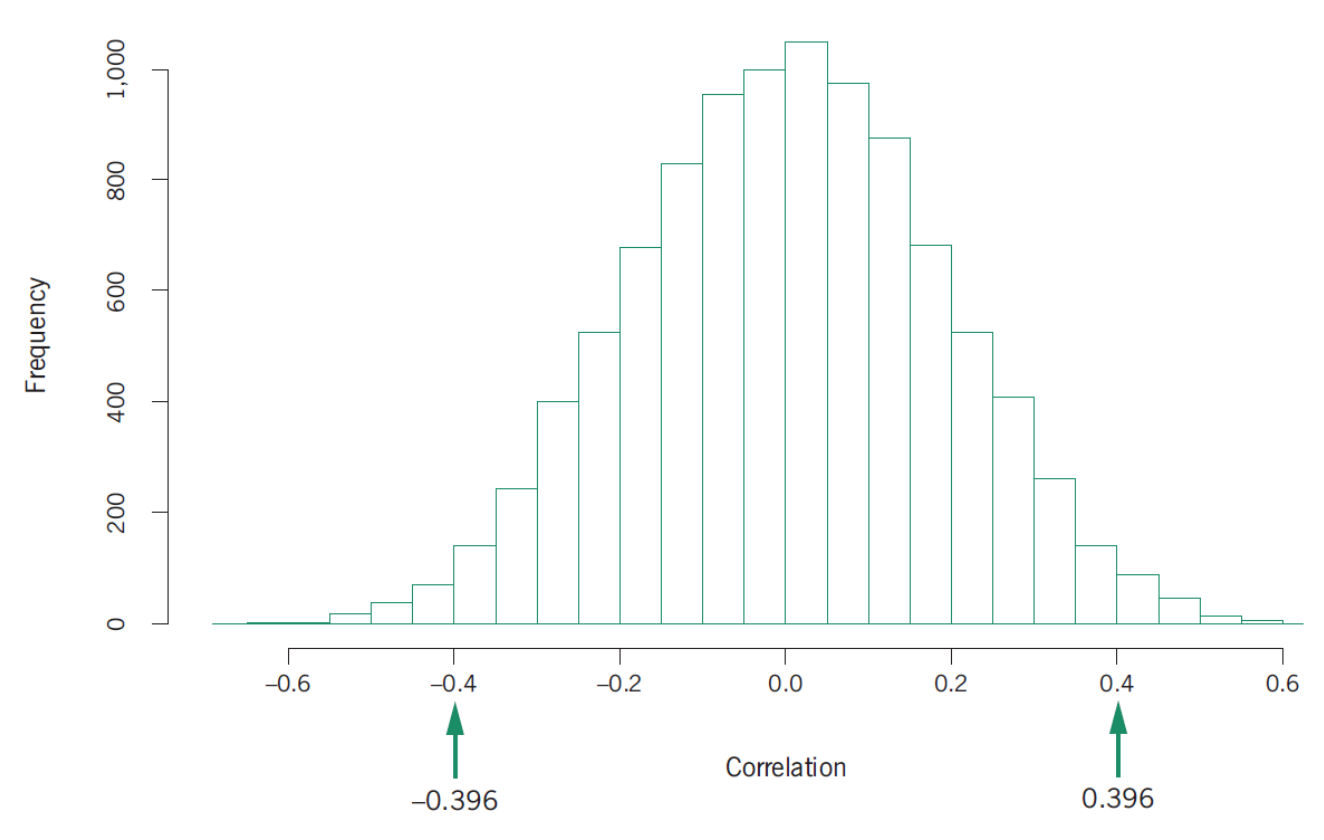
Step 4) Generate the permutation distribution



Greenacre & Primicerio 17.3

Indirect Comparison: Permutation

Step 5) Compare the observed test statistic to the permutation distribution to calculate a *P*-value



Greenacre & Primicerio 17.3

Indirect Comparison: Permutation

Advantages:

- Non-parametric: no assumptions about distribution of data
- Flexible across many test statistics
- Can be used for small sample sizes

Indirect Comparison: Permutation

Advantages:

- Non-parametric: no assumptions about distribution of data
- Flexible across many test statistics
- Can be used for small sample sizes

Disadvantage:

- May be computationally intensive

Indirect Comparison: Multi-response Permutation Procedure

Indirect Comparison: Multi-response Permutation Procedure

Multi-response permutation procedure (MRPP) is a **non-parametric** statistical test used to compare groups by testing whether within-group distances are significantly smaller than what would be expected by random chance.

Indirect Comparison: Multi-response Permutation Procedure

Multi-response permutation procedure (MRPP) is a **non-parametric** statistical test used to compare groups by testing whether within-group distances are significantly smaller than what would be expected by random chance.

Tests for significant differences among pre-defined groups (restricted to one-way tests)

Indirect Comparison: Multi-response Permutation Procedure

Multi-response permutation procedure (MRPP) is a **non-parametric** statistical test used to compare groups by testing whether within-group distances are significantly smaller than what would be expected by random chance.

MRPP is performed on the **distance** or **association matrix**.

Indirect Comparison: Multi-response Permutation Procedure

Step 1) Calculate within-group distances

Indirect Comparison: Multi-response Permutation Procedure

Step 1) Calculate within-group distances

Step 2) Compute the test statistic δ (average within-group distance)

Indirect Comparison: Multi-response Permutation Procedure

Step 1) Calculate within-group distances

Step 2) Compute the test statistic δ (average within-group distance)

Step 3) Permute (shuffle) group labels

Indirect Comparison: Multi-response Permutation Procedure

Step 1) Calculate within-group distances

Step 2) Compute the test statistic δ (average within-group distance)

Step 3) Permute (shuffle) group labels

Step 4) Generate a distribution of δ under the null hypothesis

Indirect Comparison: Multi-response Permutation Procedure

Step 1) Calculate within-group distances

Step 2) Compute the test statistic δ (average within-group distance)

Step 3) Permute (shuffle) group labels

Step 4) Generate a distribution of δ under the null hypothesis

Step 5) Calculate the *P*-value

Indirect Comparison: Multi-response Permutation Procedure

Site ID	Max Depth (m)	Gradient (%)	Elevation (m)	Canopy (%)	Herb (%)
Silvies-11	0.45	0.3	1439	0.0	55.1
Silvies-34	0.78	1.1	1487	0.0	0.0
Silvies-02	0.71	0.4	1372	29.6	0.0
Silvies-15	0.40	0.2	1471	41.1	0.0
Silvies-07	0.50	1.3	1547	52.3	0.0
Silvies-08	0.40	0.6	1492	51.4	0.0
Silvies-22	0.42	0.9	1555	54.7	0.0
Silvies-18	0.42	0.5	1510	46.2	0.0
Silvies-12	0.52	3.2	1658	51.9	0.0
Silvies-21	0.18	2.4	1713	37.5	0.0
Silvies-05	0.45	5.5	1565	46.7	0.0
Silvies-03	0.20	3.3	1634	59.0	0.0

Indirect Comparison: Multi-response Permutation Procedure

Class means and counts:

	1	2	3
delta	2.085	0.7666	2.084
n	3	5	4

Chance corrected **within-group agreement A**: 0.4108
Based on **observed delta** 1.535 and **expected delta** 2.606

Significance of delta: 0.001
Permutation: free
Number of permutations: 999

Indirect Comparison: Multi-response Permutation Procedure

Advantages:

- Non-parametric: no need to meet the assumptions of normality or homogeneity of variance
- Can be used on any distance metric

Indirect Comparison: Multi-response Permutation Procedure

Advantages:

- Non-parametric: no need to meet the assumptions of normality or homogeneity of variance
- Can be used on any distance metric; however, it is sensitive to the distance metric used!

Indirect Comparison: Multi-response Permutation Procedure

Advantages:

- Non-parametric: no need to meet the assumptions of normality or homogeneity of variance
- Can be used on any distance metric; however, it is sensitive to the distance metric used!

Disadvantages:

- Computationally intensive
- Assumes equal group sizes

Conclusion: Summary of Key Points

- **Principal Coordinates Regression (PCR)** is an indirect comparison method that uses linear regression to relate principal coordinate(s) (the explanatory variable) to the response variable(s)
- **Linear Discriminant Analysis (LDA)** is a direct comparison method that is used to classify group membership
- **Permutation** tests for statistical significance among groups/relationships when data do not meet the assumptions of normality, linearity, or homogeneity of variance
- **Multi-response Permutation Procedure (MRPP)** is a non-parametric technique used to assess the significance of group differences

We'll go over more of these next week!

Questions?