Lecture 17 - Multivariate STATS

Bill Perry

# 1.

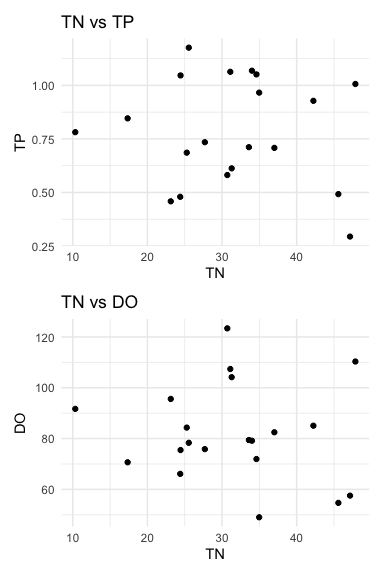
# 2. Introduction to Multivariate Statistics

## 2.1 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

# 3. Multivariate Data

* Multiple variables recorded about each object (individual, quadrat, site, 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



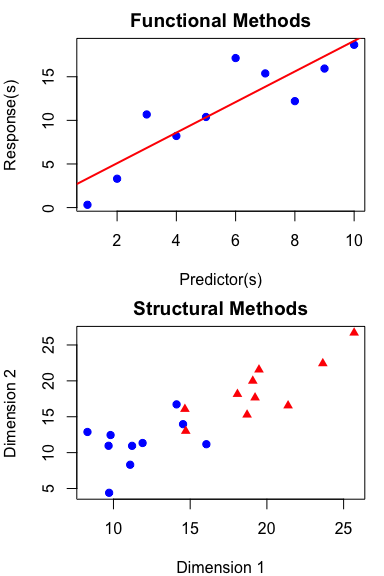
|  |
| --- |
| Multivariate data vs. multivariate analysis |
| We’ve already seen multivariate data in multiple regression and multi-factor ANOVA, but now we’ll look at cases with multiple response variables. |

# 4. Multivariate Statistics in Ecology

## 4.1 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

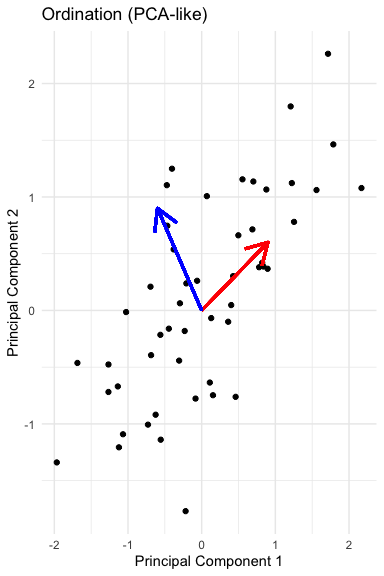


# 5. Structural Methods in Multivariate Analysis

## 5.1 Two Main Approaches

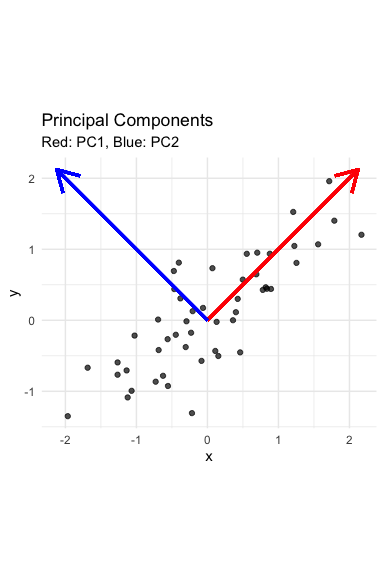
**Scaling/Ordination Methods**: - Reduce dimensions with new derived variables - Summarize patterns in data - Examples: PCA, CCA

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



# 6. Eigenvectors, Eigenvalues, and Components

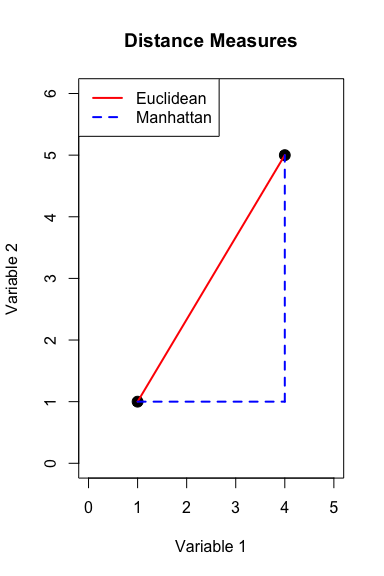
* Goal: derive new variables (principal components) that explain variation in data
* Components are linear combinations of original variables:
  + zik = c1yi1 + c2yi2 + … + cpyip
* Properties of derived variables:
  + First component explains most variation
  + Second explains most remaining variation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



|  |
| --- |
| 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. |

# 7. Distance and Dissimilarity Measures

* 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

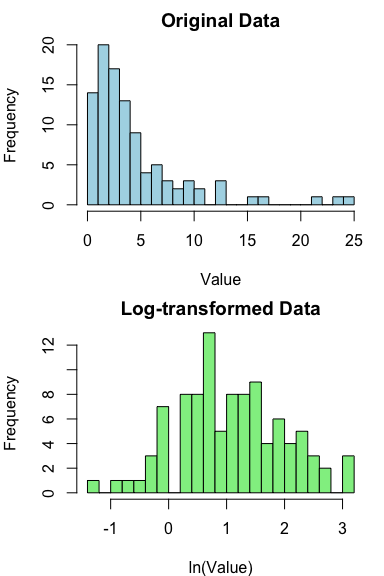


# 8. Data Transformations & Standardization

## 8.1 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

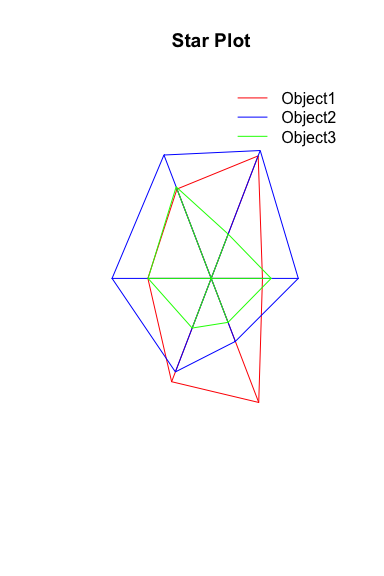


|  |
| --- |
| Why standardize? |
| Standardization ensures all variables contribute equally to the analysis regardless of their original units or scales of measurement. Without it, variables with larger values or variances would dominate the results. |

# 9. Multivariate Graphics

## 9.1 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

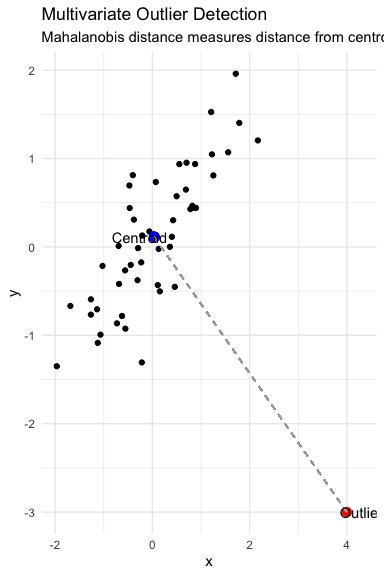


# 10. Screening Multivariate Data

## 10.1 Key Issues to Check

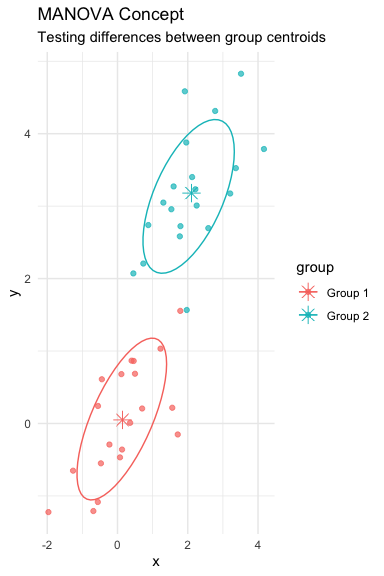
**Multivariate Outliers**: - Objects with unusual patterns across variables - Detected with Mahalanobis distance (d²) - Test against χ² distribution with p df

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



# 11. MANOVA (Multivariate Analysis of Variance)

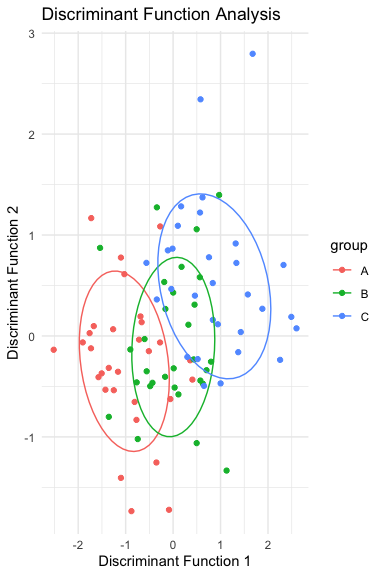
* Multivariate extension of ANOVA
* Tests for differences in group centroids based on multiple response variables
* Advantages over multiple ANOVAs:
  + Controls family-wise error rate
  + Accounts for correlations between variables
  + More powerful when variables are correlated
* Common test statistics: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



|  |
| --- |
| MANOVA Assumptions |
| * Multivariate normality * Homogeneity of variance-covariance matrices * No extreme multivariate outliers * Independence of observations |

# 12. Discriminant Function Analysis

* Mathematically similar to MANOVA
* Used for:
  + Testing differences between groups (like MANOVA)
  + Identifying variables that separate groups
  + Classifying observations into groups
* Creates linear combinations (discriminant functions) that maximize between-group differences
* Can assess how well classification performs
* Jackknifed classification provides more realistic success rates



# 13. Summary

## 13.1 Key Concepts

1. **Multivariate data** requires special techniques to account for correlations between variables
2. **Functional methods** (MANOVA) test hypotheses about group differences
3. **Structural methods** (PCA, NMDS) find patterns in data
4. **Distance measures** quantify similarities between objects
5. **Data standardization** is crucial for variables with different units
6. **Multivariate graphics** help visualize complex relationships

