Statistical methods

Clustering

Clustering is useful for getting a feel as to the structure of your data and the relationships among samples. Clustering is based on distances among objects and must therefore be paired with an appropriate distance metric.

Complete linkage (furthest neighbor)

- Also called furthest neighbor clustering
- Hierarchical method
- Algorithm:
    - Each object begins as its own cluster
    - At each step, the two clusters separated by the shortest distance are combined
    - The distance between clusters is measured as the distance between the members of each cluster that are **furthest** from one another

Single linkage (nearest neighbor)

- Also called nearest neighbor clustering
- Hierarchical method
- Algorithm:
    - Each object begins as its own cluster
    - At each step, the two clusters separated by the shortest distance are combined
    - The distance between clusters is measured as the distance between the members of each cluster that are **nearest** one another

Average linkage

- Hierarchical method
- Algorithm:
    - Each object begins as its own cluster
    - At each step, the two clusters separated by the shortest distance are combined
    - The distance between clusters is measured as the **average** distance between members of two clusters

Ward's

- Hierarchical method

- Clusters to be merged are determined by Ward's minimum variance criterion, which minimizes the total within-cluster variance

K-means

- Places objects into *k* clusters
- Minimizes within-cluster variance

Ordination

Picture your data as a big, multidimensional blob. You probably can't picture more than 3 dimensions, because what does more than three dimensions even look like? But, your data has many more than three dimensions. Within the multidimensional blob, samples are oriented to each other based on their similarity to one another. In each of these dimensions, the relationships among samples can change, because they may be affected similarly by some factors and differently by others. But how can you possibly be expected to be able to figure out which of those dimensions and relationships are most important if you can't even figure out how to picture the big blob of data?

Ordination, also referred to as gradient analysis, is a data reduction technique. It can be used to help figure out which dimensions of your data are the most important, displaying them as axes and plotting samples (or another unit of interest) along them. The first axis extracted from your data is the direction along which the variance among the data is greatest. The second axis corresponds to the second strongest gradient of variation in the data, provided it is orthogonal (unrelated) to the first. And so on and so forth until all axes have been computed.

Unconstrained ordination methods (e.g. PCoA, NMDS, PCA) seek to describe the structure of data without outside influence, and are part of exploratory analysis. Constrained ordination methods take into account (i.e. are constrained by) external data (e.g. environmental data) and can be used in hypothesis testing.

Some ordination methods are based on distances among objects, and must therefore be paired with a distance metric. Other methods are based on eigenvalues that are calculated directly from a matrix of values.

*PCoA (**P**rincipal **co**ordinates **a**nalysis)*

- Unconstrained, distance-based method
- Prone to the arch-effect
- Dissimilarities are preserved and solution is unique

- Euclidean representation of objects, but using any dissimilarity metric desired
    - I.e. maximizes linear correlation between the distance measure used and the distance within the ordination
    - Equivalent to PCA when distance is euclidean

*NMDS (**N**on-**m**etric **d**imensional **s**caling)*

- Unconstrained, distance-based method
- Desired number of axes is specified in advance, and an iterative process is used to reposition the objects in that number of dimensions so as to minimize the stress function
    - Dissimilarities are distorted and not preserved during this process
    - The starting "solution" is a random ordination and may affect the outcome
    - NMDS solutions change with the number of axes
- NMDS solutions themselves do not maximize variation along axes, but a PCA rotation is often performed on the results to maximize the variance of the solution
- Less prone to the arch-effect

*PCA (**P**rincipal **c**omponents **a**nalysis)*

- Unconstrained, eigenanalysis-based method
- Can be used on abundance data after transformation
- Detects linear relationships and preserves Euclidean distances among objects
- Takes your blob of data and rotates it until the axis of maximum variability is visible

*RDA (**R**e**d**undancy **a**nalysis)*

- Constrained
- Combination of multiple regression and PCA
    - Multivariate multiple linear regression followed by a PCA performed on the matrix of fitted values
- Can test significance using permutations
- Can be used with distance metrics or performed directly on transformed abundance values
    - If Euclidean distance is used, results are the same as performing RDA on a community abundance matrix

*CCA (**C**anonical **c**orrespondence **a**nalysis)*

- Constrained
- Very similar to RDA

- Combination of multiple regression and canonical analysis (CA)
    - Whereas PCA preserves Euclidean distance among sites, CA preserves chi-square distance
- Not as suitable for microbiome studies, because it is best served when rare species are well-sampled

Univariate statistical tests

(Binary) Logistic regression

- Uses a logistic function to model a dependent binary variable
- Often used as a predictive tool
- Assumptions
    - Dependent variable is binary
    - Observations/samples are independent
    - Little or no multicollinearity among independent variables
    - Independent variables are linearly related to the log odds

Linear regression

- Models the relationship between a scalar response and one or more explanatory variables
-
- Assumptions
    - Linear relationship between response and explanatory variables
    - Multivariate normality
    - Little or no multicollinearity among explanatory variables
    - Samples are independent
    - Homoscedasticity (equal variance)

Student's T-test

- Used to compare the means of two sample groups
- $\mathbf{H_0}$ = There is no difference in means between the two groups
- $\mathbf{H_A}$ = The two groups have different means
- Assumptions
    - Data is normally distributed in each group
    - The two groups have equal variance
    - Samples are independent or paired
- Non-parametric counterpart: Mann-Whitney U Test

Multivariate statistical tests

ANOVA

- **An**alysis **o**f **Va**riance
- Used to compare the means of three or more sample groups
- $\mathbf{H_0}$ = There is no difference in means among the groups
- $\mathbf{H_A}$ = The mean of at least one group is different
- Assumptions
    - Data is normally distributed in each group
    - The groups have equal variance
    - Samples are independent
- Non-parametric counterpart: Kruskal-Wallis test

perMANOVA

- **Per**mutational **M**ultivariate **An**alysis **o**f **Va**riance
- $\mathbf{H_0}$ = The centroids of the groups are the same for all groups
- Assumptions
    - Within-group dispersion is similar between groups
- perMANOVA tests are highly robust. As long as the groups you are comparing are of equal size, you're generally good to go
- Parametric counterpart: MANOVA
    - MANOVA is rarely used in ecology because data very rarely meets all of the assumptions

Mantel test

- A test of correlation between two matrices.
- $\mathbf{H_0}$ = Distances between objects in the response matrix are not linearly correlated with the distances between objects in the explanatory matrix.
- Can be used with multiple correlation metrics (e.g. Pearson, Spearman, Kendall)
- Beware the Mantel test! A mantel test seems simple, especially because it spits out a nice correlation index. But, it's harder to interpret than it seems. Remember that a Mantel testing is testing for correlation of *distances*. A Mantel test should only be used if both the explanatory and response variables are things you would normally express as distances (e.g. geographic distance).

ANOSIM

- **An**alysis **o**f **Sim**ilarity
- $\mathbf{H_0}$ = the average of the ranks of within-group distances is greater than or equal to the average of the ranks of between-group distances
- Assumptions

    - The ranges of (ranked) dissimilarities within groups are equal, or at least very similar
    - Can produce unreliable results if group dispersions are not similar

An important note:

perMANOVA, Mantel, and ANOSIM tests are often used interchangeably, but they are actually testing different null hypotheses. For details, see Anderson and Walsh 2013.

Redundancy analysis

- See entry in Ordination

Differential abundance

ALDEx2

- **A**NOVA-**L**ike **D**ifferential **Ex**pression
- **H_0** = There is no difference in absolute abundance of a feature between two groups
- Identifies features as differential if simple random sampling cannot explain the difference in abundance

ANCOM

- **An**alysis of **Com**positions
- **H_0** = There is no difference in absolute abundance of a feature between two groups
- Assumptions
    - Less than 25% of features (e.g. ASVs, OTUs, species) change in abundance between groups