PLS 206 Ordination Lesson Outline:

1. Moving from univariate response variables to multivariate responses
   1. In community ecology (and many other fields), the *composition* of a sample is often of interest -- usually, both the number and types of different species.
   2. Oftentimes, we like to use simple methods like tabulating summary statistics such as species richness, diversity, or total abundance to construct univariate models.
   3. However, we’re often also interested in *what* species change between samples. This data is necessarily fairly complex. For a given sampling unit, observations may consist of 100s of species, each with their own responses to different environmental variables.
2. Example species matrix
3. What is ordination?
   1. *Ordination* is a method by which a multidimensional data set (dimension = species) is attempted to be expressed in a smaller number of dimensions.
   2. Ordination comes in two flavors;
      1. Unconstrained
         1. What are the patterns in my data? How do observations distribute themselves in ordination space (usually 2 dimensions for visualizations)? Are there correlations among certain groups of sites / species?
         2. ***Exploratory*** - Hypothesis generating
      2. Constrained
         1. How do species distribute themselves across measured environmental gradients? How much of the total variation in species can I capture with environmental variables?
         2. ***Testing associations*** - Hypothesis testing
4. Ordination relies on the assumption that species distributions can be expressed over a small number of environmental gradients; we don’t need to measure a different variable for each species.
   1. E.g. Species abundances can be largely captured by knowing the pH, salinity, and soil depth of a given environment.
   2. Species gradients graph
5. In ordination, there is an emphasis on distance metrics -- how do we determine how similar two samples are from one another?
   1. Some commonly used distance metrics, and their associated methods:
      1. Euclidean distance (PCA, RDA)
      2. Chi-square distance (NMDS, CCA)
      3. Bray-Curtis distance (NMDS, db-RDA)
      4. Jaccard distance (NMDS, db-RDA)
   2. Key decision points:
      1. How does my chosen distance metric handle data with a large number of zeroes (common in community data)?
      2. Is my metric sensitive to the *total* abundance of species in a sample? Is this difference biologically meaningful, or am I more interested in relative abundance?
6. Code to walk through in class:
   1. Dune data
   2. Using the vegan package
   3. Creating a distance matrix object with vegdist()
      1. Pairwise distances between points -- there are *Nchoose2* possible pairwise combinations that can be made + the diagonal. Distance objects in R are special version of matrices, but can be coerced into standard matrices or dataframes using as.matrix() and as.dataframe().
   4. Plotting communities in an NMDS
      1. The NMDS object
         1. Pull out actual fitted points in two dimensions
         2. Add a convex hull around a specific group
      2. Stress values
         1. < .3 -- problematic fit; 2-dimensional object inaccurately represents greater than 30% of the rank-order distances between different samples
         2. .2 - .3 -- “reasonable” fit
         3. .1 - .2 -- “good” fit
         4. <.1 -- Excellent representation of the *S* dimensional object in two dimensions.
            1. In my personal experience, stress between .1 and .3 is fairly common for biological data. Stress tends to increase as the number of datapoints increases (greater number of pairwise comparisons that must be accounted for).
      3. Diagnostic fit plots
      4. **Key points to note:**
         1. Unlike a PCA, the axes themselves do not matter in an NMDS; the NMDS algorithm attempts to best describe the true distances between objects in a two-dimensional space.
            1. The emphasis here is on the rank-order distances between different samples, not necessarily the linear distance.
         2. NMDS can be rotated to suit visualization needs.
         3. NMDS visualizations will not always be constant! Different runs of the NMDS algorithm may produce different fits, particularly if there is no clear “best fit”. To keep fits consistent, start from the same random variable in R (set.seed()) and increase the number of iterations in the function call.
   5. Adding species scores
      1. Fitting and interpreting species scores
         1. Species scores are scaled by their association between different axes. A long arrow with a given species label indicates a strong “pull” of a given species in that direction.
   6. Fitting environmental variables to unconstrained ordination
      1. Adding site scores
         1. Similarly to species scores, the length of the arrow indicates the strength of association that a distribution of points has with a given environmental variable.
   7. Evaluating the significance of environmental drivers with RDA
      1. ANOVA of an RDA
         1. Inertia as a counterpart to variance
         2. Inertia captured by explanatory variables
      2. P-values of different environmental variables added
         1. Permutation test of residuals
      3. RDA visualization
         1. Importance of different canonical axes
         2. ***Note:*** This plot is different from an unconstrained ordination figure! It does not show the total variation in the dataset, but what can be captured with a given set of environmental variables.
      4. Extracting relationships between species and different environmental variables.
   8. Permutational Multivariate ANOVA (PerMANOVA; If time allows)
      1. Increasingly popular technique to compare communities that are grouped, but do not follow assumptions of multivariate normality OR are evaluated using a non-euclidean distance
      2. Permutational ANOVA evaluates the significance of 1 or more grouping (factor) variables.
         1. Calculate the total inertia of the dataset
         2. Calculate the residual within-group inertia when adding the grouping factor.
         3. Swap grouping variable, re-calculate.
         4. P - value = proportion of random suffles that reduce this ratio of captured / total variance more than the “true” arrangement.
      3. Assumptions:
         1. Variance between groups is equal. Check with the *Permutation Test for Multivariate Dispersion* (permDisp()) in vegan.