Among Site (Beta) Diversity

Z620: Quantitative Biodiversity, Indiana University

February 6, 2015

OVERVIEW

In this excercise, we move beyond the investgation of within-site α-diversity. We will explore β-diversity, which is defined as the diveristy that occurs among sites. This requires that we examine the compositional similarity of assemblages that vary in space or time.

After completing this exercise you will know how to:

1. formally quantify β-diversity
2. visualize β-diversity with heatmaps, cluster analysis, and ordination
3. test hypotheses about β-diversity using multivariate statistics

1) SETUP

A. Retrieve and Set Your Working Directory

```
rm(list = ls())
getwd()
setwd("~/GitHub/QuantitativeBiodiversity/Assignments/Beta")
```

B. Load Packages

We will be using the `vegan` package again; let's load it now.

```
require("vegan")
```

2) LOADING DATA

A. Description of Data Set

To date, we have analyzed biodiversity data sets for freshwater zooplankton, tropical trees, and soil bacteria. In this exercise, we introduce a new dataset containing information on stream fish assemblages from the Doubs river, which runs near the France-Switzerland boarder in the Jura Mountains. The data set (`doubs`) includes fish abundances, environmental variables, and spatial cooridinates for 30 sites. The data set has previously been used to demonstrate that fish communities can be good indicators of ecological zones in rivers and streams.

Let's load the `ade` package, which contains the `doubs` data set.

```
require("ade4")
data(doubs)
```

B. Introduction to a New Data Structure: Lists

While working in R this semester, we have learned about vectors, matrices, and data frames. Here we introduce another data structure: a **list**. In R, a list is an object that contains a collection of other objects of similar or different types.

C. Exploring the Doubs River Data Set

We can use the `str()` function to describe the attributes of `doubs`, which is a list. Because this dataset is somewhat complex, we can pass the "max.level = 1" argument to minimize the `str()` output. Also, you can use the the dollar sign (`$`) between the list name (`doubs`) and objects within the list (e.g., `env`) to explore the data set. Last, recall that you can use `help()` to learn more about a dataset contained in a package.

```
str(doubs, max.level = 1)
```

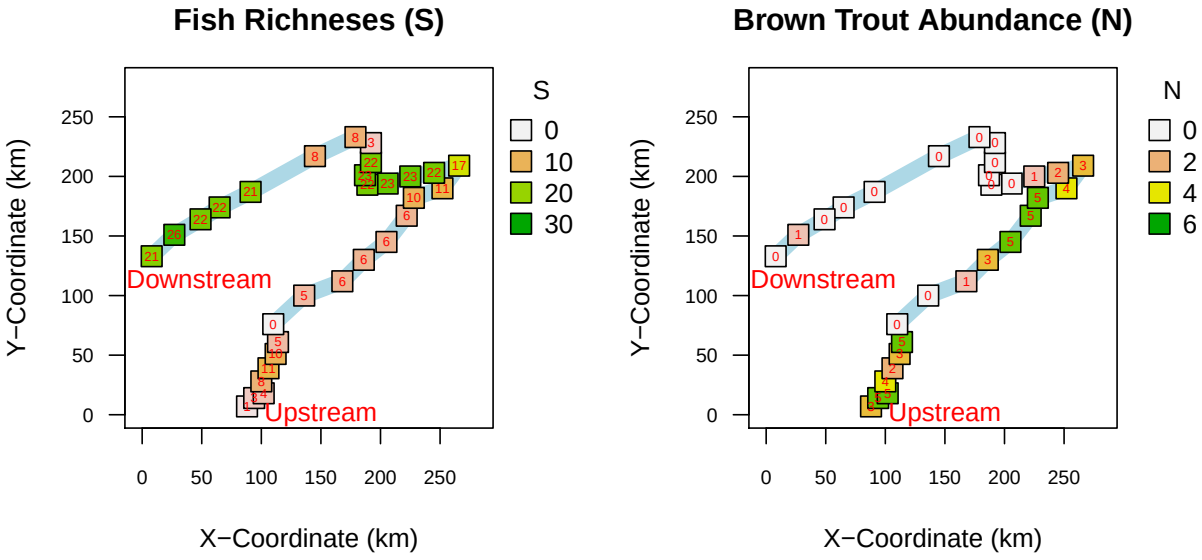
Question 1: Describe some of the attributes of the `doubs` dataset.

a. How many objects are in `doubs`?
b. What types of data structures are contained in `doubs`?
c. What are the units of nitrate ("nit") in the stream water?
d. How many fish species are there in the `doubs` data set?

Answer 1a:
Answer 1b:
Answer 1c:
Answer 1d:

D. Visualizing the Doubs River Dataset

There is a wealth of information in the `doubs` dataset that can be used to address various issues related to β-diversity. For example, we might use the environmental or spatial data to develop or test a hypothesis. Below we have generated two plots of the `doubs` fish data. The first plot shows fish richness at each site in the stream. The second plot shows the abundance of a particular fish species, Brown Trout (*Salmo trutta*), at each site in the stream.



Question 2: Answer the following questions based on the spatial patterns of richness (i.e., α-diversity) and Brown Trout (*Salmo trutta*) abundance in the Doubs River.

a. How does fish richness vary along the sampled reach of the Doubs River?
b. How does Brown Trout (*Salmo trutta*) abundance vary along the sampled reach of the Doubs River?
c. What do these patterns say about the limitations of using richness when examining patterns of biodiversity?

Answer 2a:
Answer 2b:
Answer 2c:

3) QUANTIFYING BETA-DIVERSITY BETWEEN TWO SAMPLES

There are various ways to quantify β-diversity. Perhaps one of the simplest metrics is **Whittaker's β-Diversity**, which was developed by Robert Whittaker (1960). This classic index is useful for comparing β-diversity between two samples. The equation for Whittaker's β-Diversity is as follows: $\beta_W = \frac{S}{\bar{\alpha}} - 1$, where S = the total number of species recorded across sites (i.e., γ-diversity) and $\bar{\alpha}$ is the mean richness (i.e., α diversity) among sites. Subtracting 1 scales β_w from 0 (minimum β-diversity) to 1 (maximum β-diversity).

We can write β_W as a function in R as follows:

```
beta.w <- function(site1 = "", site2 = ""){
  site1 = subset(site1, select = site1 > 0)               # Removes absences
  site2 = subset(site2, select = site2 > 0)               # Removes absences
  gamma = union(colnames(site1), colnames(site2))         # Gamma species pool
  s     = length(gamma)                                   # Gamma richness
  a.bar = mean(c(specnumber(site1), specnumber(site2)))   # Mean sample richness
  b.w   = round(s/a.bar - 1, 3)
  return(b.w)
}
```

Question 3: Using the `beta.w` function above, answer the following questions:

a. What is the β-diversity for fish assemblages sampled from site 1 and site 2 of the `doubs` data set?
b. Based on the formula for β_w, what does this value represent?

Answer 3a:
Answer 3b:

4) QUANTIFYING BETA-DIVERSITY FOR TWO OR MORE SAMPLES

Often, we often want to compare the diversity of more than just a pair of samples. For example, it would be nice to be able to compare fish assemblages for *all* of sites in the Doubs River. In this section we will estimate β-diversity for multiple samples. During this process, you will learn how to generate similarity and dissimilarity matrices for different data sets that will be needed for visualizing and quantifying β-diversity.

A. Introducing the Resemblance Matrix

In order to quantify β-diversity for more than two samples, we need to introduce our second primary ecological data structure: the **Resemblance Matrix**. In the context of biodiversity, a resemblance matrix is a data structure that calculates the pairwise **similarity** or **dissimiliarity** for all samples in a site-by-species matrix. The resemblance matrix can be generated from a site-by-species matrix containing incidence (presence-absence) data or abundance data. In the sections below, we describe some of the similarity and dissimilarity metrics that are commonly used for constructing a resemblance matrix. Throughout this handout, we adopt the notations of Legendre & Legendre (2012); this book can be electronically accessed via the IU library (see the course syllabus [http://goo.gl/y4oK7c]).

B. Incidence-Based Measures of Similarity and Dissimilarity

When you are working with presence-absence data, you can use the following metrics for generating a similarity or dissimilarity matrix.

Index	Equation	Properties
Jaccard	$S_7 = \frac{a}{a+b+c}$	Compares the number of shared species to the number of species in the combined assemblages (global view)
Sørensen	$S_8 = \frac{2a}{(2a+b+c)}$	Compares the number of shared species to the mean number of species in a single assemblage (local view)

In the above table, a = the number of species shared between assemblages, b = the number of unique species in the first assemblage, and c = the number of unique species in the second assemblage.

Question 4: Answer the following questions about incidence-based measures of similarity:

a. What are the differences between Jaccard and Sørensen metrics?
b. When might you use one instead of the other?
c. In what situations would these metrics of β-diversity fail?

Answer 4a:
Answer 4b:
Answer 4c:

Notes on incidence-based similarity: Jacard and Sørensen are perhaps the two most commonly used incidence-based measures of similarity. Others include Ochiai, Kulczynski-Cody, and Lennon, which can be found in Table 6.1 of Magurran & McGill (2011). The differences in these measures include how means are calculated (Sørensen = harmonic mean, Ochiai = geometric mean, and Kulczynski-Cody = arithmetic mean), and how unique species are dealt with if only one sample has unique species (Lennon). Also, it is important to note that these metrics calculate similarity, but can (and in many cases should) be converted to dissimilarity (D). In `vegan`, disimilarities (D) are usually returned from functions instead of similarities (S). The conversion between similarity and dissimilarity is calculated as `D = 1 - S`.

C. Abundance-Based Measures of Similarity and Dissimilarity

When you are working with abundance data, you can use the following metrics for generating a similarity or dissimilarity matrix.

Index	Equation	Properties
Bray-Curtis Dissimilarity	$D_{14} = \frac{\sum_{j=1}^{p} \lvert y_{1j} - y_{2j} \rvert}{\sum_{j=1}^{p} (y_{1j} + y_{2j})}$	A quantitative version of the Sørensen index. Commonly used measure of similarity. Also known as the *percentage difference*.
Morisita-Horn	$S_{MH} = \frac{2 \sum_{j=1}^{p} y_{1j} \cdot y_{2j}}{\left(\sum_{j=1}^{p} y_{1j}^2 + \sum_{j=1}^{p} y_{2j}^2 \right)}$	A measure of *compositional overlap*. Uses squared differences in relative abundance and thus is influenced by abundant species. Resistant to undersampling.

In the above table y_{1j} is the abundance of each species (1:p) in site 1 and y_{2j} is the abundance of each species (1:p) in site 2. As with incidence-based measures, there are many other options for calculating similarity or dissimilarity between communities, including Mean Character Difference, Canberra, Coefficient of Divergence, and Gower.

D. A Cautionary Note on Other Measures of Distance

There are other distance measures that you are likely to encounter because they are widely used. It is important to note that some of these should *not* be used for abundance-based data in biodiversity analyses. This is because they lead to the **Species Abundance Paradox**, which is a phenomenon that occurs when the distance between two sites that have no species in common is smaller than the distance between sites with shared species. In particular, the paradox described above arises when Euclidean Distance and Manhattan Distance are used:

Euclidean Distance: $D_1 = \sqrt{\sum_{j=1}^{p} (y_{1j} - y_{2j})^2}$

Manhattan Distance: $D_7 = \sum_{j=1}^{p} |y_{ij} - y_{2j}|$

Methods exist to transform species abundance data prior to implementing metrics like Euclidean Distance and Manhattan Distance (e.g., Chord transformation and Chi-Square transformation). It is recommended that you read more about these distance metrics if you are interested in implementing them to address your research questions.

E. Constructing the Resemblance Matrix

Conveniently, `vegan` includes many of the similarity metrics used to construct a resemblance matrix. These metrics can be calcualted using the `vegdist()` function. Let's use `vegdist` to create a resemblance matrix for the fish assemblages in the Doubs River. Before that, we'll need to remove site 8 from `doubs` because for some reason it has no observations.

```
fish <- doubs$fish
fish <- fish[-8, ] # Remove site 8 from data

# Calculate Jaccard
fish.dj <- vegdist(fish, method = "jaccard", binary = TRUE)

# Calculate Bray-Curtis
fish.db <- vegdist(fish, method = "bray")
```

Now that we've created a resemblance matrix, it would be nice to visualize the fish assemblages of the Doubs River. As a start, we can print the Bray-Curtis-based resemblance matrix in the console:

```
fish.db
```

From this, you will see a large diagonal matrix. Typically, resemblance matrices just show the upper or lower triangle of values. This is because the two triangles have the same information. However, you can generate a square resemblance matrix with the following command:

```
fish.db <- vegdist(fish, method = "bray", upper = TRUE, diag = TRUE)
```

Question 5: Does the resemblance matrix (`fish.db`) represent similarity or dissimilarity? What information in the resemblance matrix led you to arrive at your answer?

Answer 5:

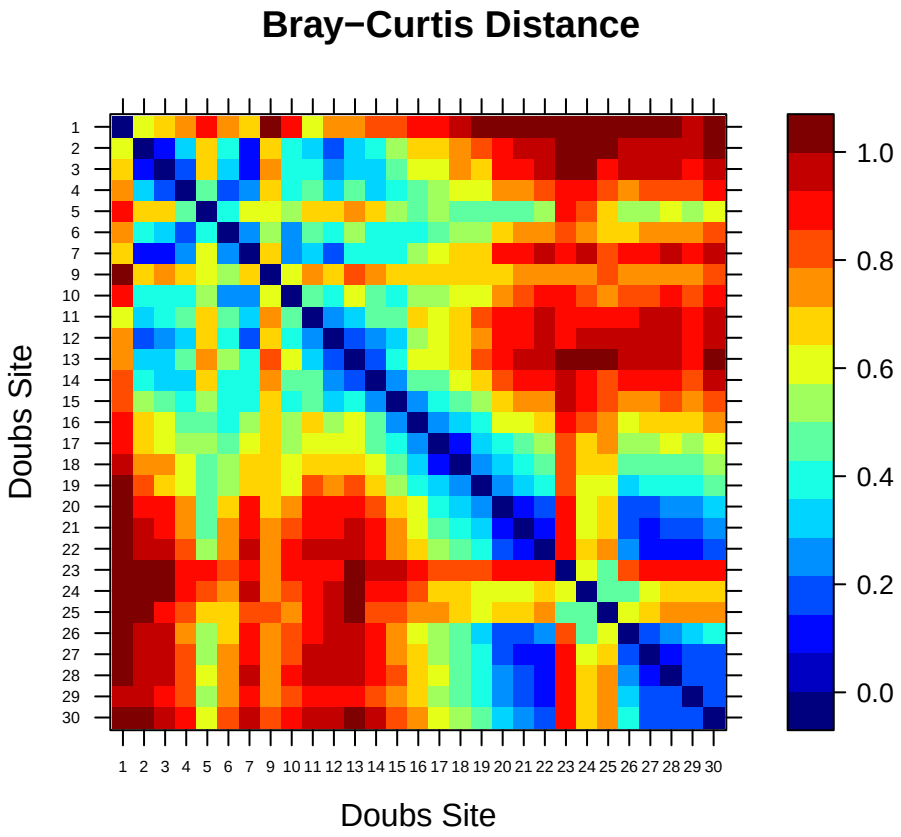
5) VISUALIZING BETA-DIVERSITY

A. Heatmaps

One way to visualize β-diversity is to plot the data in our resemblance matrix using a **heatmap**. Heatmaps are a two-dimensional, color representation of a data matrix. Here we are going to use the `levelplot()` function in the `lattice` package of R. This function will allow us to make a basic heatmap of our resemblance matrix. First, there are a few things we need to do:

1. Define a color palette. R includes many predefined color palettes; however, we are going to make our own color palette.
2. Ensure that our resemblance matrix is plotted correctly. In particular, we need specify the order in which we want our sites to be plotted in the heatmap.

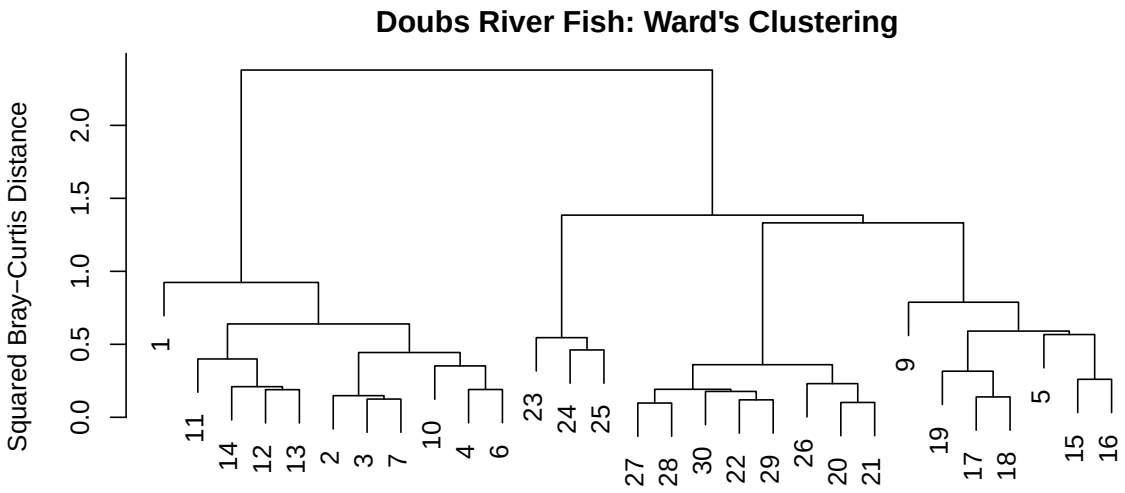
```
# Custome Color Palette
jet.colors <- colorRampPalette(c("#00007F", "blue", "#007FFF", "cyan",
                                 "#7FFF7F", "yellow", "#FF7F00", "red",
                                 "#7F0000"))
# Define Order of Sites
order <- rev(attr(fish.db, "Labels"))

# Plot Heatmap
levelplot(as.matrix(fish.db)[, order], aspect = "iso", col.regions = jet.colors,
          xlab = "Doubs Site", ylab = "Doubs Site", scales = list(cex = 0.5),
          main = "Bray-Curtis Distance")
```



B. Cluster Analysis

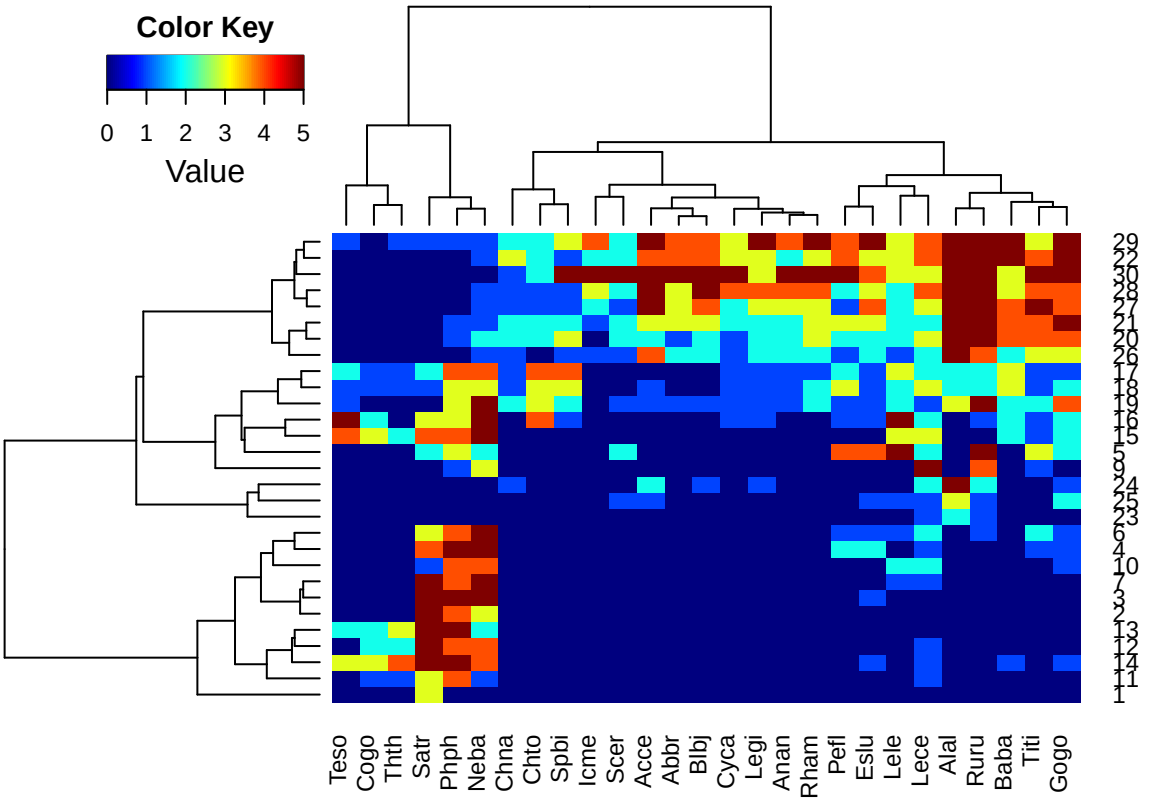
Another common way to visualize β-diversity is through cluster analysis. Cluster analysis is an exploratory technique that assigns objects to groups based on their similarity to one another. In this exercise, we will use hierarchical clustering, specifically **Ward's Clustering**. Ward's Clustering (a.k.a., Ward's minimum variance method) is an agglomerative clustering technique based on the linear model criterion of least squares. The method minimizes within-cluster sums-of-squared distances between sites. However, there are numerous methods for clustering (e.g., Single Linkage, UPGMA, UPGMC), which can influence the conclusions that you draw from your analysis. See chapter 8 of Legendre and Legendre (2012) for more information on the various clustering methods that can be used.

```
# Perform Cluster Analysis
fish.ward <- hclust(fish.db, method = "ward.D2")

# Plot Cluster
par(mar = c(1, 5, 2, 2) + 0.1)
plot(fish.ward, main = "Doubs River Fish: Ward's Clustering",
     ylab = "Squared Bray-Curtis Distance")
```



Another clustering tool that aids in visualization for exploratory purposes is `heatmap.2()`, which is a function in the `gplots` package. This tool generates a cluster diagram that allows one to examine the abundance of different fish species in different sites.

```
require(gplots)
heatmap.2(as.matrix(fish), distfun = function(x) vegdist(x, method = "bray"),
          hclustfun = function(x) hclust(x, method = "ward.D2"),
          col = jet.colors(100), trace = "none", density.info = "none")
```



Question 6: Based on cluster analyses and the introdutory plots that we generated after loading the data, develop a hypothesis for the `doubs` data set?

Answer 6:

C. Ordination

The primary aim of ordination is to represent multiple objects in a reduced number of orthogonal (i.e., independent) axes. The first axis of an ordination explains the most variation in the data set, followed by the second axis, then the third, and so on, where the total number of axes is less than or equal to the number of objects. Ordination plots are particularly useful for visualizing the similarity among objects. For example, in the context of β diversity, sites that are closer in ordination space have species assemblages that are more similar to one another than sites that are further apart in ordination space.

There are various ordination techniques that can be applied to multivariate biodiversity data. Common methods include: Principal Components Analysis (PCA), Correspondence Analysis (CA), Principal Coordinates Analysis (PCoA), Factor Analysis (FA), and Nonmetric Multidemensional Scaling (NMDS). When choosing an ordination technique, careful consideration should be given to the data type (continuous vs. categorical), model assumptions, and the underlying mathematical procedures that are invovled.

i. An Overview of Principal Coordinates Analysis (PCoA)

In this exercise, we focus on Principal Coordinates Analysis (PCoA), which is sometimes referred to as metric multidimensional scaling. PCoA starts with creating a matrix, $\boldsymbol{A}$, which is a transformed and centered version of distance matrix, $\boldsymbol{D}$. Because these steps preserve all distances, PCoA is a flexible ordination technique that allows us to use virtually any distance metric (e.g., Jaccard, Bray-Curtis, Gower, Euclidean, etc.). The dimensionality of $\boldsymbol{A}$ is then reduced by determining each eigenvector, $\boldsymbol{u}_i$, and eigenvalue, $\boldsymbol{\lambda}_i$, that solve the following equation: $\boldsymbol{A}\boldsymbol{u}_i = \boldsymbol{\lambda}_i\boldsymbol{u}_i$. Finally, each eigenvector, $\boldsymbol{u}_i$, is scaled to length $\sqrt{\boldsymbol{\lambda}_i}$ to obtain the principal coordinates.

To conduct a PCoA, we will use the `cmdscale()` function from the `stats` package, which performs PCoA. The input for this function is our resemblance matrix (Bray-Curtis distances for the `doubs` dataset). In addition, we are going to set $k = 3$ (number of dimensions we want returned) and *eig* = TRUE (which saves the eigenvalues).

```
fish.pcoa <- cmdscale(fish.db, eig = TRUE, k = 3)
```

ii. Interpetting PCoA Output

The `cmdscale` function produces a list of output. The first item *points* contains the coordinates for each site in each reduced dimension. The second item *eig* contains the eigenvalues. The last three items pertain to other options of the analysis that we will not cover here.

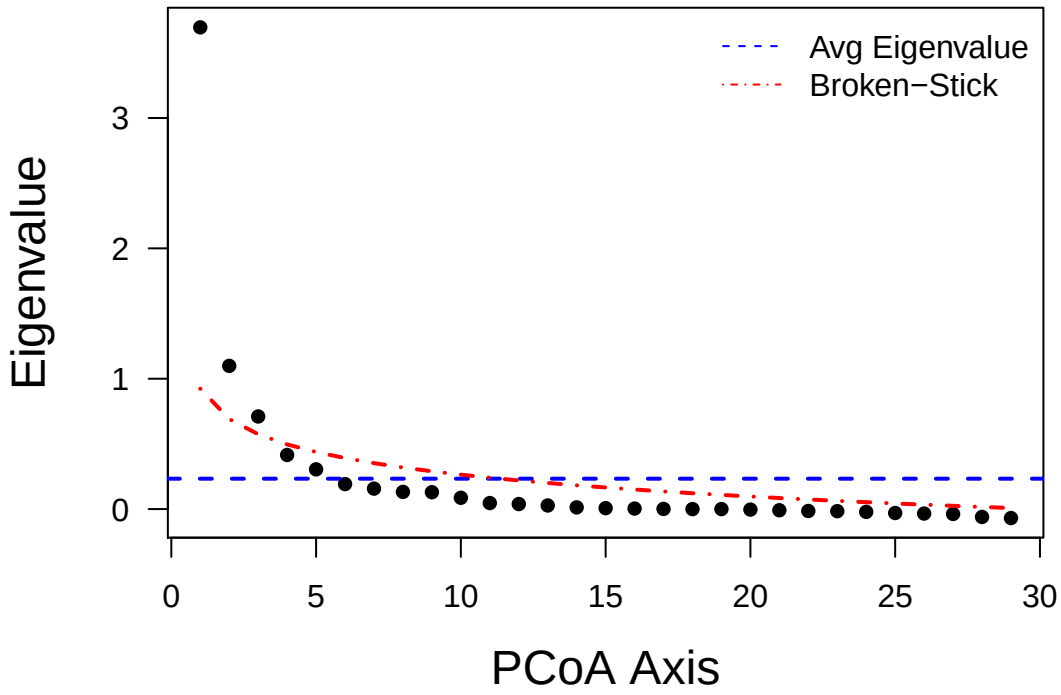
First, we want to examine the eigenvalues. The eigenvalues are the scaling factors that allowed us to reduce the dimensionality of a data set. The eigenvalues can also be used to calcualte the amount of variation that is explained by each orthogonal axis. To do this, we divide the eigenvalue of each axis by the sum of all eigenvalues. In the following chunk of R code, we quantify the percent variation in the `doubs` data set that is explained by the first three axes of the PCoA.

```
explainvar1 <- round(fish.pcoa$eig[1] / sum(fish.pcoa$eig), 3) * 100
explainvar2 <- round(fish.pcoa$eig[2] / sum(fish.pcoa$eig), 3) * 100
explainvar3 <- round(fish.pcoa$eig[3] / sum(fish.pcoa$eig), 3) * 100
sum.eig <- sum(explainvar1, explainvar2, explainvar3)
```

Another way to evaluate our analysis is to assess whether or not the first few PCoA axes capture a disproportionately large amount of the total explained variation. First, the eigenvalues associated with the first few axes should be larger than the average of all the eigenvalues (*Kaiser-Guttman criterion*). Second, we can compare the eigenvalues associated with the first few axes to the expectations of the *broken-stick model*, which we introduced in the α-diversity exercise when discussing species abundance distributions (SAD). In the current context, the broken stick model assumes that the total sum of eigenvalues decreases sequentially with ordered PCoA axes.

We will evaluate these two criteria with the following plots:

```
# Define Plot Parameters
par(mar = c(5, 5, 1, 2) + 0.1)

# Plot Eigenvalues
plot(fish.pcoa$eig, xlab = "PCoA Axis", ylab = "Eigenvalue",
     las = 1, cex.lab = 1.5, pch = 16)

# Add Expectation based on Kaiser-Guttman criterion and Broken Stick Model
abline(h = mean(fish.pcoa$eig), lty = 2, lwd = 2, col = "blue")
b.stick <- bstick(29, sum(fish.pcoa$eig))
lines(1:29, b.stick, type = "l", lty = 4, lwd = 2, col = "red")

# Add Legend
legend("topright", legend = c("Avg Eigenvalue", "Broken-Stick"),
       lty = c(2, 4), bty = "n", col = c("blue", "red"))
```



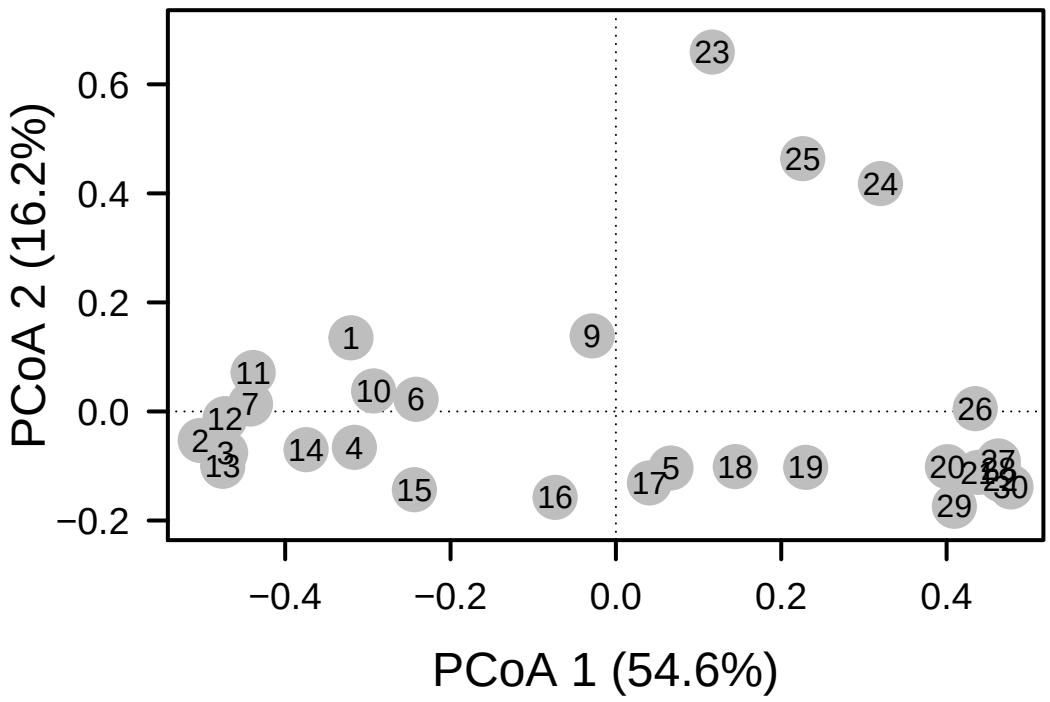
Question 7: Based on the three criteria described above, does the PCoA do a good job of explaining variation in the `doubs` data set? Please justify.

Answer 7:

iii. Creating a PCoA Ordination Plot

Having evaluated the PCoA output, now we will create an ordination plot. We will plot all of the fish assemblages of the Doubs River for the first two PCoA axes.

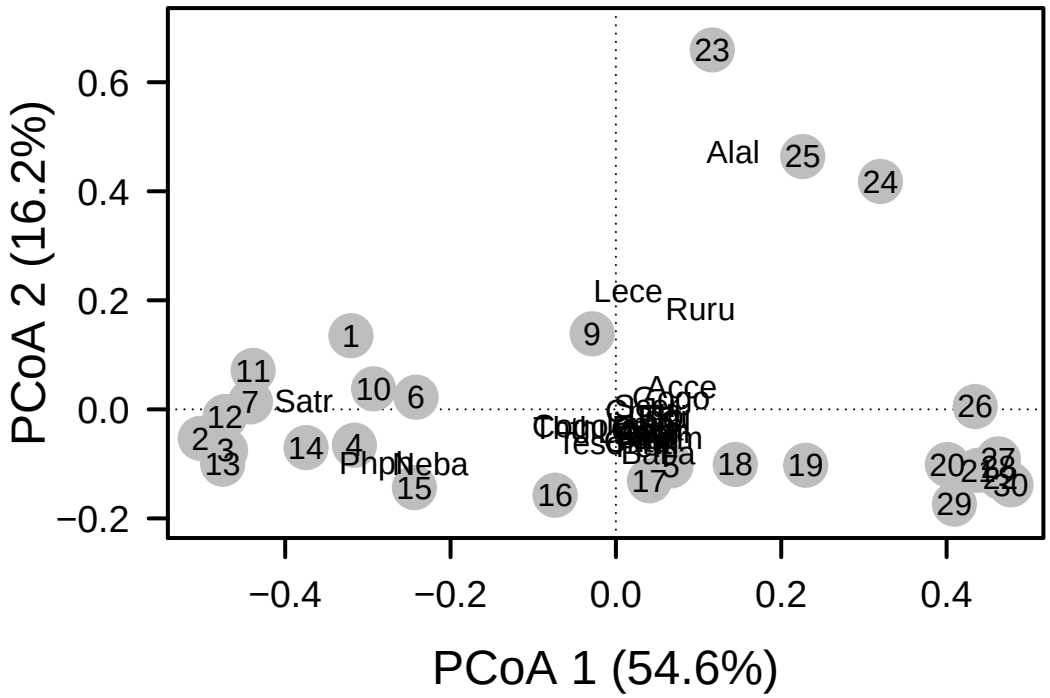
```
# Define Plot Parameters
par(mar = c(5, 5, 1, 2) + 0.1)

# Initiate Plot
plot(fish.pcoa$points[ ,1], fish.pcoa$points[ ,2], ylim = c(-0.2, 0.7),
     xlab = paste("PCoA 1 (", explainvar1, "%)", sep = ""),
     ylab = paste("PCoA 2 (", explainvar2, "%)", sep = ""),
     pch = 16, cex = 2.0, type = "n", cex.lab = 1.5, cex.axis = 1.2, axes = FALSE)

# Add Axes
axis(side = 1, labels = T, lwd.ticks = 2, cex.axis = 1.2, las = 1)
axis(side = 2, labels = T, lwd.ticks = 2, cex.axis = 1.2, las = 1)
abline(h = 0, v = 0, lty = 3)
box(lwd = 2)

# Add Points & Labels
points(fish.pcoa$points[ ,1], fish.pcoa$points[ ,2],
       pch = 19, cex = 3, bg = "gray", col = "gray")
text(fish.pcoa$points[ ,1], fish.pcoa$points[ ,2],
     labels = row.names(fish.pcoa$points))
```



iv. Identifying and Visualizing Influential Species in PCoA

Basic ordination plots allow us to see how samples separate from one another. A logical follow-up is to ask what features of the data set are driving the observed divergence among points. In the Doubs River example, sites are separating along the PCoA axes owing to variation in the abundance of different fish species. We can get a better sense of "who" is contributing to this trend by plotting explanatory vectors (i.e., species coordinates) in ordination space. We can obtain this information using the `add.spec.scores()` function in the `BiodiversityR` package. These coordinates reflect the strength and direction that each species has on the ordination of the different sites.

```
require("BiodiversityR")

# Calculating Relative Abundance
fishREL <- fish
  for(i in 1:nrow(fish)){
    fishREL[i, ] = fish[i, ] / sum(fish[i, ])
    }

# Calculate and Add Species Scores
fish.pcoa <- add.spec.scores(fish.pcoa,fishREL,method = "pcoa.scores")
text(fish.pcoa$cproj[ ,1], fish.pcoa$cproj[ ,2],
     labels = row.names(fish.pcoa$cproj), col = "black")
```



A more quantitative way of identifying influential species involves determining the correlation of each species along the PCoA axes. To do this, we will use the `add.spec.scores` function again. Then we can identify a correlation-coefficient cutoff (e.g. $r = 0.70$) to pull out important species. Finally, we will use the `envfit()` function from the `vegan` pacakge, to conduct a permutation test on these correlations.

```r
spe.corr <- add.spec.scores(fish.pcoa, fishREL, method = "cor.scores")$cproj
corrcut  <- 0.7       # user defined cutoff
imp.spp  <- spe.corr[abs(spe.corr[, 1]) >= corrcut | abs(spe.corr[, 2]) >= corrcut, ]

# Permutation Test for Species Abundances Across Axes
fit <- envfit(fish.pcoa, fishREL, perm = 999)
```

Question 9: Address the following questions about the ordination results of the `doubs` data set:

a. Generate a hypothesis about the grouping of sites in the Doubs River based on fish community composition.
b. Generate a hypothesis about which fish species are potential indicators of river quality.
c. Do the different approaches described in the ordination section agree or disagree? Explain.

Answer 9a:
Answer 9b:
Answer 9c:

6) HYPOTHESIS TESTING

The visualization tools that we just learned about (i.e., heatmaps, cluser analysis, and ordination) are powerful for exploratory analysis and for *generating* hypotheses. In this section we introduce some methods that are better suited for *testing* hypotheses and predictions related to β-diversity.

A. Multivariate Procedures for Categorical Designs

PERMANOVA stands for permutational multivariate analysis of variance (Anderson 2001). It is a multivariate analog to univariate ANOVA and has less restrictions than parametric multivariate analysis of variance (MANOVA). As the name suggests, it evaluates differences according to a specified model by randomly permuting the data. PERMANOVA can easily handle simple designs, but also accomodates nested and higher-order studies. In addition, it can deal with missing values and unbalanced designs. The PERMANOVA output is similar to the output of classic ANOVA; it includes (pseudo) F-tests, p-values, and R^2 values. We will implement PERMANOVA with the `adonis()` function in the `vegan` package.

To run `adonis`, you first need a factor vector or matrix that specifies your treatments and replicates. This vector or matrix can either be provided in the form of a file that you open in R (e.g. `read.table()`) or a user-specified vector or matrix that you write within R.

Earlier work done in the Doubs River suggested that the river can be broken into four distinct regions based on fish habitat quality. The first region (sites 1-14) has been rated as being "high quality". The second (sites 15 - 19) and fourth (sites 26 - 30) regions have been rated as being "moderate quality". And the third region (sites 20 - 25) has been rated has being "low quality".

Even though the Doubs River is not a controlled experiment, let's test the hypothesis that fish community composition is affected by river quality rankings.

```r
# Create "Factors"" vector
quality <- c(rep("HQ", 13), rep("MQ", 5), rep("LQ", 6), rep("MQ", 5))

# Run PERMANOVA with adonis function
adonis(fish ~ quality, method = "bray", permutations = 999)
```

Question 10: Based on the PERMANOVA results, evaluate the prediction that river quality influences fish community composition.

Answer 10:

B. Mutivariate Procedures for Continuous Designs

i. Mantel Test

A Mantel test is essentially a multivariate correlation analysis. It produces an r value that is analagous to the Pearson's correlation coefficient. In addition, it produces a p-value that is derived from the deviation of observed correlation to that of correlations derived from randomizations of the two matrices.

In the following section, we will peform a Mantel test using the `mantel()` function in `vegan`. This requires that we first have two distance matrices to compare. Here, we will compare the Bray-Curtis distance matrix we created earlier with a new distance matrix of environmental factors (doubs$env). After creating the distance matrices, we will test the hypothesis that fish assemblages are correlated with stream environmental variables.

```
# Define Matrices
fish.dist <- vegdist(doubs$fish[-8, ], method = "bray")
env.dist <- vegdist(scale(doubs$env[-8,]),method = "euclid")

#Mantel Test
mantel(fish.dist,env.dist)
```

Question 11: What do the results from our Mantel test suggest about fish diversity and stream environmental conditions? How might this relate to your hypothesis about stream quality influencing fish communities?

Answer 11:

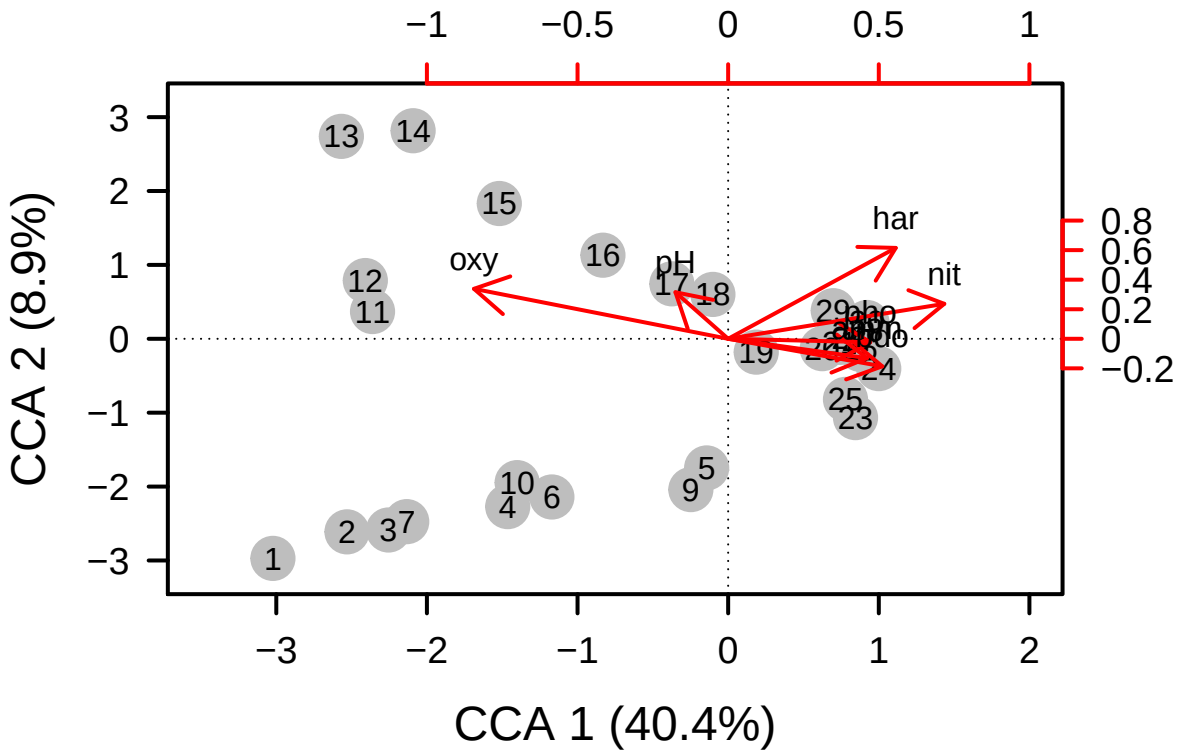
ii. Constrained Ordination

Another way we can test hypotheses with continuous data is to use **constrained ordination**, which is sometimes referred to as canonical ordination. Constrained ordination explores the relationships between two matrices: an **explanatory matrix** and a **response matrix**. Canonical correspondence analysis (CCA) and redundancy analysis (RDA) are two types of constrained ordination. These techniques are based on the linear model framework and thus can be used to formally test hypotheses. Constrained ordination works by first conducting multivariate multiple linear regression followed either by correspondence analysis (CA) with CCA or Principal Components Analysis (PCA) with RDA, while using the matrix of fitted values to obtain a constrained ordination. A permutation test can then be used to test for overall significance.

Here, we will use environmental data to conduct a CCA on the fish assemblages of the Doubs River. We will start by creating an explanatory matrix that contains water chemistry data. We will then use the `cca()` function from the `vegan` package. Note, we have to specify that we want the `cca` function in the `vegan` package because there are `cca` functions in both `vegan` and `ade4`! We will then use permutation tests to evaluate the significance of our model. Finally, we will test the influence of each environmental variable on the constrained axes.

```
# Define Environmental Matrix
env.chem <- as.matrix(doubs$env[-8 , 5:11])

# Conduct CCA
```

```r
doubs.cca <- vegan::cca(fish ~ env.chem)

# Permutation Tests
anova(doubs.cca, by = "axis")
cca.fit <- envfit(doubs.cca, env.chem, perm = 999)
cca.fit

# Calculate Explained Variation
cca.explainvar1 <- round(doubs.cca$CCA$eig[1] /
                         sum(c(doubs.cca$CCA$eig, doubs.cca$CA$eig)), 3) * 100
cca.explainvar2 <- round(doubs.cca$CCA$eig[2] /
                         sum(c(doubs.cca$CCA$eig, doubs.cca$CA$eig)), 3) * 100

# Define Plot Parameters
par(mar = c(5, 5, 4, 4) + 0.1)

# Initiate Plot
plot(scores(doubs.cca, display = "wa"), xlim = c(-3.5, 2), ylim = c(-3.2, 3.2),
     xlab = paste("CCA 1 (", cca.explainvar1, "%)", sep = ""),
     ylab = paste("CCA 2 (", cca.explainvar2, "%)", sep = ""),
     pch = 16, cex = 2.0, type = "n", cex.lab = 1.5, cex.axis = 1.2, axes = FALSE)

# Add Axes
axis(side = 1, labels = T, lwd.ticks = 2, cex.axis = 1.2, las = 1)
axis(side = 2, labels = T, lwd.ticks = 2, cex.axis = 1.2, las = 1)
abline(h = 0, v = 0, lty = 3)
box(lwd = 2)

# Add Points & Labels
points(scores(doubs.cca, display = "wa"),
       pch = 19, cex = 3, bg = "gray", col = "gray")
text(scores(doubs.cca, display = "wa"),
     labels = row.names(scores(doubs.cca, display = "wa")))

# Add Environmental Vectors
vectors <- scores(doubs.cca, display = "bp")
row.names(vectors) <- c("pH", "har", "pho", "nit", "amm", "oxy", "bdo")
arrows(0, 0, vectors[,1] * 2, vectors[, 2] * 2,
       lwd = 2, lty = 1, length = 0.2, col = "red")
text(vectors[,1] * 2, vectors[, 2] * 2, pos = 3,
     labels = row.names(vectors))
axis(side = 3, lwd.ticks=2, cex.axis=1.2, las = 1, col = "red", lwd = 2.2,
     at = pretty(range(vectors[, 1])) * 2, labels = pretty(range(vectors[, 1])))
axis(side = 4, lwd.ticks=2, cex.axis=1.2, las = 1, col = "red", lwd = 2.2,
     at = pretty(range(vectors[, 2])) * 2, labels = pretty(range(vectors[, 2])))
```



Question 12: Based on the CCA, what are the environmental variables that seem to be contributing to stream water quality for fish assemblages?

Answer 12:

7) HOMEWORK

1. We are going to revsit the soil bacteria data set that we introduced during the the α-diversity exercise. You may recall that 16S rRNA sequences were generated from replicate sites from four different land-use treatments (T1 = agriculture, T7 = grassland, DF = deciduous forest, and CF = coniferous forest). On top of this, we characterized bacteria composition from the different soils under experimentally manipulated dry vs. wet condtions. Thus, we have a 2 x 2 full-factorial design (with the exception of a missing sample from the DF land-use treatment). More background on this study can be found in Aanderud et al. 2015 (http://goo.gl/TRgISq)

Two files are available to you in the "~/QB2015_[username]/Beta/data"" directory:

a. Site-by-species matrix ("soilbacfull.txt")
b. A factor file that describes the treatments ("soil.factors.txt")

Use a combination of visualization and hypothesis-testing techniques described above to interpret the results from the exepriment in a β-diversity framework. Peform this analysis using an incidence-based distance metric and an abundance-based distance metric. Compare and contrast the outcomes.

2. Use Knitr to create a pdf of your completed alpha_exercise.Rmd document, push it to GitHub, and create a pull request. The due date for this assignment will be announced in class and/or canvas.