Using R for community ecology analyses
EDAMAME
Kellogg Biological Station
June 27, 2015

Objectives of exercise:
-visualize and test hypotheses with multivariate data
-develop tool kit for application to data types discussed previously in class

After this exercise you should be able to do the following:

- Load data tables into the R Statistical Computing environment
- Manipulate data matrices
- Transform and standardize community composition data sets
- Calculate pairwise distance (resemblance) matrices summarizing relationships between community composition or "species" occurrence patterns
- Visualize community composition data using ordination and clustering techniques
- Test hypotheses about differences in community composition across samples or sites
- Identify correlations between community composition and environmental conditions of sites

Overview of lab activities:

1) Brief introduction to matrices and R
2) Loading data, data transformations, and visualization with a "toy" matrix
3) Hypothesis testing using analysis of similarity (ANOSIM) and KBS LTER microbial community data
4) Incorporating environmental information using correlation and Mantel Tests
5) Summary discussion and questions

1) Brief introduction to matrices and R

a) Matrices

-A matrix is an ordered set of numbers listed in rectangular form (think Excel sheet!).

$$A = \begin{matrix} 2 & 3 & 4 & 4 \\ 6 & 2 & 1 & 9 \\ 9 & 5 & 2 & 8 \end{matrix}$$

-The matrix above (matrix A) has 3 rows and 4 columns, it is a 3 x 4 (read as "three by four") matrix.
-A matrix with a single row or a single column is often called a "vector".
-The elements of a matrix are referred to by their row number and column number (always in that order). From the matrix above, $A_{3,2} = 5$.
-(micro)ecological data, especially community composition data, often takes the form of a matrix or vector.

b) R Statistics Package

-R is an open source, command prompt-based software that allows loading and manipulation of data in matrix form
-Individuals are able to contribute **packages** to the Comprehensive R Archive Network (CRAN). Packages provide useful statistical, mathematical, and graphical processes (**functions**) in an easy to access form
-R and its packages are easy to download and install from the R Project for Statistical Computing website (http://www.r-project.org)
-A matrix is one type of **data structure** in R; others include: vector, list, and data frame; see additional data types handout
-To manipulate matrices in R, we have to use the previously described matrix notation.
-Each object (and therefore matrix) in R has a name (e.g. "A", "data", or "stuart")
-The same row by column format is used to access portions of a matrix

>To access a whole matrix, use: A
>To access a single element, use: A[1,3]
>To access a whole row, use: A[2,]
>To access a whole column, use: A[,4]

-The basic package of R and any additional packages you install include a series of functions.
-Functions are called by typing their name followed by an open and close parenthese, e.g. plot(); However, the function can't do what it is designed to do without giving it object(s) and rules to work with. These pieces of information go within the parentheses and are called **arguments**
-Typing help(plot) or ?plot will give you the help file for the function plot; within the help file is a list of arguments that must be passed to the function

plot and what defaults are used if an argument of a given type isn't passed to the function

Final Note: There is a fairly steep learning curve with R, but it is an extremely powerful tool. There are numerous books and tutorials available on the web. I believe the best way to become proficient is to commit to using R for a project even if it seems like it will take a really long time to do something you know how to do in some other program (Excel, Primer, etc.). Don't get frustrated and use the web and others as a resource!

2) Loading data, data transformations, and visualization

-Upon startup R Studio should open with some information about R and the version of R installed on your machine
-R only "sees" the files in the working directory. To access information from a file, you must put that file in your current working directory or change your current working directory to the directory containing the file.
-To change your current working directory, use the "Files" tab. Click "…" in the upper right to navigate to your working directory. Then under "More" select "Set as working directory"
-Alternatively from the command prompt you can use the setwd() function, but you must provide the full path as a string as an argument.

****Installation of the vegan package: We will rely heavily on a multivariate statistics package called "vegan". To install packages in R Studio use the "Packages" tab. Click "Install in the upper left and search for the desired package. Be sure to check the box "Install dependencies".

***Note: Any line in this document that begins with the R line prompt, ">", is intended for you to run at the line prompt on your machine, **but I encourage you to keep your work in a well-commented script!** Expected output from R will be printed in smaller, italicized text in this document.

a) Loading a data table

The function read.table() is used to read in delimited text files. Often R will assign the data type "data frame" to loaded text files, but we want it to be of the type "matrix" so we use the function as.matrix() along with read.table(). Don't forget to assign the matrix a name so you can continue to use it (e.g. "data").

```
>data=as.matrix(read.table("Table1.txt",header=TRUE,sep="\t",row.names=
1))
```

```
> data
      Sample1 Sample2 Sample3 Sample4
OTU1     25      80      60      90
OTU2     25       0       5       0
OTU3     25       0       5       0
OTU4     25       0       0       0
OTU5      0      20      30      10
> dim(data)
[1] 5 4
> data[1,]
   Sample1 Sample2 Sample3 Sample4
OTU1     25      80      60      90
> data[,3]
[1] 60  5  5  0 30
```

Questions

1) What does the output of the function dim() report?

2) For which OTU did we recover the most sequences in our study?

3) An ordination is a visual depiction of complex data often found in ecology. Usually these are two-dimensional scatter plots where the pairwise distances between points represent the multivariate, compositional distance between samples. Draw an ordination of the four samples in this dataset.

b) Data transformations

Often community composition data is transformed from raw abundance to presence-absence or relative abundance. We can also easily calculate sample richness using the presence-absence matrix.

```
> dataPA=(data>0)*1
> dataPA
      Sample1 Sample2 Sample3 Sample4
OTU1        1       1       1       1
OTU2        1       0       1       0
OTU3        1       0       1       0
OTU4        1       0       0       0
OTU5        0       1       1       1
> rich=colSums(dataPA)
> rich
Sample1 Sample2 Sample3 Sample4
      4       2       4       2
```

There is a slower and faster way to calculate relative recovery/abundance from raw data. The faster approach requires something called a "for loop". In a "for loop", a process is repeated for each member of a list. In our case each column *i* of the data matrix is divided by the sum of that column *i*.

Slower method

```
> dataREL=matrix(0,5,4)
> dataREL[,1]=data[,1]/sum(data[,1])
> dataREL[,2]=data[,2]/sum(data[,2])
> dataREL[,3]=data[,3]/sum(data[,3])
> dataREL[,4]=data[,4]/sum(data[,4])
> dataREL
      Sample1 Sample2 Sample3 Sample4
OTU1   0.25    0.8    0.60    0.9
OTU2   0.25    0.0    0.05    0.0
OTU3   0.25    0.0    0.05    0.0
OTU4   0.25    0.0    0.00    0.0
OTU5   0.00    0.2    0.30    0.1
> colSums(dataREL)
Sample1 Sample2 Sample3 Sample4
      1       1       1       1
```

Faster method

```
> dataREL2=data
> for(i in 1:4){dataREL2[,i]=data[,i]/sum(data[,i])}
> dataREL2
      Sample1 Sample2 Sample3 Sample4
OTU1   0.25    0.8    0.60    0.9
OTU2   0.25    0.0    0.05    0.0
OTU3   0.25    0.0    0.05    0.0
OTU4   0.25    0.0    0.00    0.0
OTU5   0.00    0.2    0.30    0.1
> colSums(dataREL2)
Sample1 Sample2 Sample3 Sample4
      1       1       1       1
```

Another tranformation that is useful at times is to transpose a matrix (switch the rows and columns). This is important because the orientation of the matrix determines whether you evaluate community composition or species occurrence patterns

```
> t(dataREL2)
        OTU1 OTU2 OTU3 OTU4 OTU5
Sample1 0.25 0.25 0.25 0.25  0.0
Sample2 0.80 0.00 0.00 0.00  0.2
Sample3 0.60 0.05 0.05 0.00  0.3
Sample4 0.90 0.00 0.00 0.00  0.1
```

The basis of many multivariate statistical approaches is a pairwise distance matrix or resemblance matrix. This is a symmetrical (reflected across the diagonal) square matrix. Sometimes only the lower triangle of this matrix is shown for brevity. Each element of the matrix is an index of distance or dissimilarity between two samples. If we calculate a distance matrix, DIST, DIST[1,1] is the dissimilarity between sample 1 and sample 1 and therefore equals 0. The diagonal is made up of all "self

comparisons and is filled with 0s. DIST[2,1] is the distance between sample one and sample two. Common dissimilarity indices include: Euclidean, Bray-Curtis, Sorenson's, and Jaccard (see Appendix 1 for calculations). A really useful function for calculating distance matrices in R is vegdist(). This is in a package contributed by Jari Oksanen called vegan. Packages can be installed using the "Packages" tab in R Studio. To load a package, the function library() is used.

```
> library(vegan)
This is vegan 2.0-10

> samplePA.dist=vegdist(t(dataPA),method="jaccard")
> samplePA.dist
        Sample1 Sample2 Sample3
Sample2     0.8
Sample3     0.4     0.5
Sample4     0.8     0.0     0.5
> otuPA.dist=vegdist(dataPA,method="jaccard")
> otuPA.dist
     OTU1 OTU2 OTU3 OTU4
OTU2 0.50
OTU3 0.50 0.00
OTU4 0.75 0.50 0.50
OTU5 0.25 0.75 0.75 1.00
> sampleREL.dist=vegdist(t(dataREL2),method="bray")
> sampleREL.dist
        Sample1 Sample2 Sample3
Sample2    0.75
Sample3    0.65    0.20
Sample4    0.75    0.10    0.30
```

Questions

1) Which of our samples had the highest α-diversity (richness)?

2) Why would it be a good idea to standardize each sample to a sum of one?

3) Were the presence-absence and relative distance matrices the same? Why or why not?

c) Visualization: Ordination, Clustering, and Heatmaps

A common ordination technique used in ecology is principle coordinates analysis (PCoA). This is a useful way to visualize complex community composition data in two dimensions. The function for PCoA in R is cmdscale(). Other common ordination techniques include Correspondence Analysis (CA; R function is cca()) and non-metric multidimensional scaling (NMMDS; R function is metaMDS()).

```
> samplePA.pcoa=cmdscale(samplePA.dist)
```

```
> samplePA.pcoa
                [,1]       [,2]
Sample1  0.4813025  0.08424920
Sample2 -0.3168243  0.02953543
Sample3  0.1523462 -0.14332005
Sample4 -0.3168243  0.02953543
> sampleREL.pcoa=cmdscale(sampleREL.dist)
> sampleREL.pcoa
              [,1]       [,2]
Sample1  0.52990579  0.024747578
Sample2 -0.22049875  0.008313729
Sample3 -0.09518054 -0.154254092
Sample4 -0.21422650  0.121192785
```

Another common approach for visualizing multivariate data is clustering. Here we will use Unweighted Pair Group Method with Arithmetic Mean (UPGMA) clustering with the R function hclust().

```
> samplePA.clust=hclust(samplePA.dist)
> sampleREL.clust=hclust(sampleREL.dist)
```

The most common function for plotting data is plot().
***Note: You can scroll through your plots with the left and right arrows in the upper left of the "Plots" tab in R Studio.

```
> plot(samplePA.pcoa[,1],samplePA.pcoa[,2],cex=0,main="PA sample PCOA")
> text(samplePA.pcoa[,1],samplePA.pcoa[,2],seq(1,4),cex=1.5)
> plot(sampleREL.pcoa[,1],sampleREL.pcoa[,2],cex=0,main="standardized
sample PCOA")
> text(sampleREL.pcoa[,1],sampleREL.pcoa[,2],seq(1,4),cex=1.5)
> plot(samplePA.clust,main="PA samples")
> samplePA.dist
          Sample1 Sample2 Sample3
Sample2    0.8
Sample3    0.4     0.5
Sample4    0.8     0.0     0.5
> plot(sampleREL.clust,main="standardized samples")
> sampleREL.dist
          Sample1 Sample2 Sample3
Sample2    0.75
Sample3    0.65    0.20
Sample4    0.75    0.10    0.30
```

Another popular method for depicting microbial community data is a heatmap. This figure type combines cluster analysis with shaded cells indicating the relative abundance of each "species" in each sample.

```
> heatmap(dataREL2,scale="none",labCol=c('S1','S2','S3','S4'))
```

Questions

1) Were the presence-absence and relative PCOA ordinations the same? Why or why not?

2) Do sample differences depicted in your cluster diagram agree with your heatmap?

3) Hypothesis testing with multivariate data (ANOSIM, MRPP, PERMANOVA)

The data we are using in this section was created using MOTHUR and KBS LTER (deciduous, MB, and standard agriculture, MA) 16S rDNA sequences. We used two operational taxonomic unit (OTU) definitions, 97% and 90% sequence identity. Although patterns can be identified using visualization techniques the significance or likelihood that the observed patterns are non-random can only be confirmed using hypothesis testing approaches, like Analysis of Similarity (ANOSIM).

a) Loading, Transforming, and Visualizing

In this section, we will repeat the methods applied to the simplified matrix used in section 2 a-c.

```
> lter=as.matrix(read.table("LTERsoil_otus.txt",sep="\t",header=TRUE))
> lter3=lter[2:302,1:6]
> lter10=lter[2:114,7:12]
> dim(lter3)
[1] 301   6
> dim(lter10)
[1] 113   6
> colSums(lter3)
MA1S1_0.03 MA1S2_0.03 MA1S3_0.03 MB3S1_0.03 MB3S2_0.03 MB3S3_0.03
88       94       91       94       91       83
> colSums(lter10)
MA1S1_0.1 MA1S2_0.1 MA1S3_0.1 MB3S1_0.1 MB3S2_0.1 MB3S3_0.1
88      94      91      94      91      83
> lter3PA=(lter3>0)*1
> lter10PA=(lter10>0)*1
> colSums(lter3PA)
MA1S1_0.03 MA1S2_0.03 MA1S3_0.03 MB3S1_0.03 MB3S2_0.03 MB3S3_0.03
     67       71       71       73       73       67
> colSums(lter10PA)
MA1S1_0.1 MA1S2_0.1 MA1S3_0.1 MB3S1_0.1 MB3S2_0.1 MB3S3_0.1
    39      45      38      39      37      38
> lter3REL=lter3
> lter10REL=lter10
> for(i in 1:6){lter3REL[,i]=lter3[,i]/sum(lter3[,i]);
lter10REL[,i]=lter10[,i]/sum(lter10[,i])}
> lter3REL.dist=vegdist(t(lter3REL),method="bray")
```

```
> lter10REL.dist=vegdist(t(lter10REL),method="bray")
> lter3REL.pcoa=cmdscale(lter3REL.dist)
> lter10REL.pcoa=cmdscale(lter10REL.dist)

> plot(lter3REL.pcoa,cex=2, pch=c(rep(15,3),rep(21,3)),main="97% OTUs
standardized samples PCOA")
> legend('topleft',c('AG','DF'),pch=c(15,21))
> plot(lter10REL.pcoa,cex=2, pch=c(rep(15,3),rep(21,3)),main="90% OTUs
standardized samples PCOA")
> legend('bottomleft',c('AG','DF'),pch=c(15,21))
> plot(hclust(lter3REL.dist),main="97% OTUs standardized samples ")
> plot(hclust(lter10REL.dist),main="90% OTUs standardized samples")

>lter3REL.ca=cca(t(lter3REL)
>lter3REL.sc=scores(lter3REL.ca,display='sites')

>plot(lter3REL.pcoa,cex=2,pch=c(rep(15,3),rep(21,3)),xlab='PCOA1',ylab='P
COA2',main="97% OTUs standardized samples PCOA")
> legend('topleft',c('AG','DF'),pch=c(15,21))
>plot(lter3REL.sc[,1],lter3REL.sc[,2],cex=2,pch=c(rep(15,3),rep(21,3)),xlab='
CA1',ylab='CA2',main="97% OTUs standardized samples CA")
> legend('topleft',c('AG','DF'),pch=c(15,21))
```

Questions

1) Were the 97% or 90% sequence identity OTU-based data richer? Why?

2) Does your OTU definition influence the conclusions you take away from your ordinations? If so, why? If not, when would you expect it to?

3) Do your PCOA and CA ordinations agree quantitatively? How about qualitatively?

b) Hypothesis testing

Analysis of similarity (ANOSIM), Multi Response Permutation Procedure (MRPP), and PERMANOVA are multivariate versions of Analysis of Variance (ANOVA). Kevin introduced ANOSIM, but here is a reminder of what ANOSIM does. It compares the distances between samples within *a priori* defined groups to distances between samples across defined groups. However the actual distances are not used, ANOSIM is non-parametric and also is based on ranks of distances rather than the actual distances. The ANOSIM statistic (R) is scaled between -1 and 1 with 1 indicating strong differences in composition. Multiple Response Permutation Procedure (MRPP) is a similar method to ANOSIM that is not rank based. For both approaches, significance of the statistic is determined by permutation of group labels. PERMANOVA is the most closely related to ANOVA of the three as it is based on a pseudo-F-distribution. PERMANOVA is implemented in R using the function adonis(). Being more closely related to ANOVA, this approach allows for incorporation of more complex model designs.

> anosim(lter3REL.dist,grouping=c(rep(1,3),rep(2,3)),permutations=1000)

Call:
anosim(dis = lter3REL.dist, grouping = c(rep(1, 3), rep(2, 3)), permutations = 1000)
Dissimilarity: bray

ANOSIM statistic R: 1
Significance: 0.08

Based on 1000 permutations

> anosim(lter10REL.dist,grouping=c(rep(1,3),rep(2,3)),permutations=1000)

Call:
anosim(dis = lter10REL.dist, grouping = c(rep(1, 3), rep(2, 3)), permutations = 1000)
Dissimilarity: bray

ANOSIM statistic R: 0.8148
Significance: 0.085

Based on 1000 permutations

>mrpp(t(lter3REL),grouping=c(rep(1,3),rep(2,3)),distance="bray",permutations=1000)

Call:
mrpp(dat = t(lter3REL), grouping = c(rep(1, 3), rep(2, 3)), permutations = 1000, distance = "bray")

Dissimilarity index: bray
Weights for groups: n

Class means and counts:
1 2
delta 0.7422 0.7574
n 3 3
Chance corrected within-group agreement A: 0.07452
Based on observed delta 0.7498 and expected delta 0.8101

Significance of delta: 0.10390
Based on 1000 permutations

>adonis(lter3REL~c(rep(1,3),rep(2,3)),method="bray",permutations=1000)
Call:
adonis(formula = lter3REL.dist ~ c(rep(1, 3), rep(2, 3)), permutations = 1000)

Df SumsOfSqs MeanSqs F.Model R2 Pr(>F)
c(rep(1, 3), rep(2, 3)) 1.00000 0.52382 0.52382 1.86227 0.3177 0.08791 .
Residuals 4.00000 1.12513 0.28128 0.6823
Total 5.00000 1.64895 1.0000
---
*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Questions
1) What was a null and alternative hypothesis for your ANOSIM?

2) Does your significance for any of the tests exactly match the ones in your handout? Why not? What about the ANOSIM R statistic? Why or why not?

3) Does your OTU definition influence the conclusion from your ANOSIM?

4) Do the three approaches generally agree? Talk with your neighbor about the differences between the three approaches. ***Use help files, books, google, faculty, and TAs to facilitate this discussion.***

4) Incorporating other information

Often we want to find correlation between two types of data. For example, what environmental features correlate strongly with a given community composition or does geographic location of sites correlate with community composition? In addition, we can identify the relative ability of geography and local environmental characteristics to explain community composition. For this section we will use EDAMAME's Centralia soils dataset. Community composition from 18 soil cores in Centralia, Pennsylvania was determined using 16S rDNA Illumina sequencing seq and QIIME 1.9.1. In addition to characterizing the bacterial community composition the chemical environment was described.

a) Loading, transforming, and visualizing In this section, we will repeat the methods applied in the previous two sections. In the second plot we will code the points by fire affectedness. These groups will also be used for running an ANOSIM.

```
#Reading in tables
>soils=read.table("otu_table_mc2_w_tax_even4711_CollapseReps_forR.tx
t",header=TRUE,sep="\t",row.names=1,stringsAsFactors=FALSE)

#remove the last column, the Consensus Lineage
>soils= soils[,-ncol(soils)]

>dim(soils)
#24364 18

>colSums(soils)
#C01  C02  C03  C04  C05  C06  C07  C08  C09  C10  C11  C12  C13
C14  #C15  C16  C17  C18
#4711 4711 4711 4711 4711 4711 4711 4711 4711 4711 4711 4711 4711
4711 #4711 4711 4711 4711

>env=read.table("Centralia_Collapsed_Map_ForR.txt",header=TRUE,sep="
\t",row.names=1, check.names=FALSE,stringsAsFactors=FALSE)
```

```
#colnames(soils)
#[1] "C01" "C02" "C03" "C04" "C05" "C06" "C07" "C08" "C09" "C10"
"C11" #"C12" "C13" "C14" "C15" "C16" "C17" "C18"

>colnames(env)
#[1] "Sample"                 "Replicate"              "Extraction"
#"collection_date"
#[5] "latitude"               "longitude"
#"AirTemperature_C"       "SoilTemperature_to10cm"
#[9] "FireFront"              "Classification"
#"DateFire_Elick2011"     "OrganicMatter_500"
#[13] "NO3N_ppm"               "NH4N-ppm"               "pH"
#"SulfateSulfur_ppm"
#[17] "K_ppm"                  "Ca_ppm"                 "Mg_ppm"
#"Fe_ppm"

#Relativize dataset
>soilsREL=soils
>for(i in 1:ncol(soils)){soilsREL[,i]=soils[,i]/sum(soils[,i])}

#Sanity check - column sums should add up to 1
>colSums(soilsREL)

#Make resembance matrix
>soilsREL.dist=vegdist(t(soilsREL),method="bray")

#Make PCoA and plot
>soilsREL.pcoa=cmdscale(soilsREL.dist)

>plot(soilsREL.pcoa,cex=0,main="standardized soil samples PCOA")
>text(soilsREL.pcoa[,1],soilsREL.pcoa[,2],rownames(soilsREL.pcoa))
>unique(env$Classification
>Class=rep('black',ncol(soils))
>Class[env$Classification=="Recovered"]='red'
>Class[env$Classification=="Reference"]='green'
>plot(soilsREL.pcoa,cex=1.5,pch=16,col=Class,main="standardized soil
samples PCOA")
>legend('topright',c('Fire
Affected','Recovered','Reference'),pch=16,col=c('black','red','green
'),box.lty=0)

#Test with ANOSIM
>anosim(soilsREL.dist,grouping=Class)

#Call:
#anosim(dat = soilsREL.dist, grouping = Class)
#Dissimilarity: bray
#
#ANOSIM statistic R: 0.3564
#Significance: 0.002
#
#Permutation: free
#Number of permutations: 999
```

Questions 1) Did we see a strong effect of fire?

b) Correlation of community composition and environmental features Often we are interested in determining whether ordination axes scores correlate with any major environmental features. This suggests that a given environmental characteristic may play a role in determining community composition. Vectors can also be plotted on our ordinations to summarize the correlation structure between samples and environmental characteristics. Species scores can also indicate potential environmental preferences by different OTUs. These methods are primarily exploratory and other approaches are needed to test hypotheses.

```
#Test for strength and significance of environmental drivers
#reduce env to only measured environmental conditions
>env2=as.matrix(env[,c(7,8,12:20)])
>EnvCor=cbind(round(cor(env2,soilsREL.pcoa[,1]),2),round(cor(env2,so
ilsREL.pcoa[,2]),2))
>colnames(EnvCor)=c("PCoA1", "PCoA2")
>EnvCor
#PCoA1 PCoA2
#AirTemperature_C         0.51  0.09
#SoilTemperature_to10cm   0.74  0.23
#OrganicMatter_500       -0.01 -0.24
#NO3N_ppm                 0.38  0.06
#NH4N-ppm                 0.37  0.02
#pH                       0.06  0.55
#SulfateSulfur_ppm        0.07  0.32
#K_ppm                    0.08  0.18
#Ca_ppm                  -0.05  0.55
#Mg_ppm                  -0.18  0.33
#Fe_ppm                   0.38 -0.16
>soilsEF=envfit(soilsREL.pcoa,env2)
>plot(soilsREL.pcoa,cex=0,main="standardized soil samples PCOA")
>text(soilsREL.pcoa[,1],soilsREL.pcoa[,2],rownames(soilsREL.pcoa))
>plot(soilsEF)
>plot(env2[,"SoilTemperature_to10cm"],soilsREL.pcoa[,1],xlab="Soil
Temperature",ylab="PCOA1")
>cor(soilsREL.pcoa[,1],env[,"SoilTemperature_to10cm"])
#[1] 0.7397104
```

Just like we can calculate the correlation between lake water color and the first axis of the PCOA, we can estimate the strength of correlation between the environmental distance or geographic distance between the lakes and the compositional distance between communities. The approach to do so is called the Mantel Test. This estimates the correlation (*r*) between two square matrices and estimates a significance of the correlation using permutation. The Mantel statistic (r) varies

between -1 and 1 with proximity to -1 or 1 indicating the strength of correlation and sign indicating direction of correlation.

```
>env.dist=vegdist(env2,method="euclidean")
>mantel(env.dist,soilsREL.dist)
#Mantel statistic based on Pearson's product-moment correlation
#Call:
#mantel(xdis = env.dist, ydis = soilsREL.dist)
#Mantel statistic r: 0.2544
#      Significance: 0.042
#Upper quantiles of permutations (null model):
#  90%   95% 97.5%   99%
#0.205 0.245 0.276 0.330
#Permutation: free
#Number of permutations: 999
```

Questions

1) Do the environmental vectors we added to our PCOA agree with the correlation table we created?

2) Do we observe an influence of local environmental effects on community composition?

3) What are the strongest and weakest local environmental correlates?

Appendix 1

Community Distance Indices for samples *j* and *k* calculated across all species *i*

$$Euclidean:\ d_{j,k} = \sqrt{\sum \left(x_{i,j} - x_{i,k}\right)^2}$$

$$Bray - Curtis:\ d_{j,k} = \frac{\sum \left| x_{i,j} - x_{i,k} \right|}{\sum \left(x_{i,j} + x_{i,k}\right)}$$

$$Sorenson's:\ Bray - Curtis, but\ with\ presence - absence\ data$$

$$Jaccard:\ \frac{2B}{1+B}, where\ B = Bray - Curtis\ distance$$

Appendix 2

INDIRECT GRADIENT ANALYSES (only species data is used to create ordinations; unconstrained ordinations)

Distance-based approaches

-Principal Coordinates Analysis (PCoA; Metric Dimensional Scaling)

Projects multidimensional distances into two or three distances. Maximizes correlation between multidimensional distances and reduced dimension distances. Eigenanalysis on distance matrix. When the distance metric used is Euclidean PCoA=PCA. Assumes a linear response to gradients.

-Nonmetric Multidimensional Scaling (NMDS/MDS)

Maintains rank order of distances rather than absolute distances. This avoids assumptions used in PCoA, but can at times find local solutions rather than global. This occurs because an iterative approach is used to minimize disagreement between the distance matrix and the ordination using the "stress" value. Also, information about species identity or relationships is lost. The likelihood of finding local solutions can be minimized by estimating the NMDS ordination from multiple starting configurations.

Eigenanalysis-based

-Linear

+Principal Components Analysis (PCA)

The simplest and oldest eigenanalysis-based method. Rotats the original data matrix onto a set of new axes, such that the maximum variance is

represented by the first axis and the second most variance is represented by an orthogonal axis. Assumes linear responses of species. Suffers from "the horseshoe effect" where the second axis is curved and/or twisted relative to the first and does not represent a true, independent secondary gradient.

-Unimodal

+Correspondence Analysis (CA; Reciprocal Averaging)

Also known as reciprocal averaging because one method for finding the solution involves repeated averaging of sample scores and species scores. CA maximizes the correspondence between species scores and sample scores. CA assumes unimodal responses of species along the axes. This allows for the potentially useful interpretation of species scores as optima along unmeasured environmental gradients. The second and higher axes can suffer from "the arch effect", which is similar to the horseshoe effect of PCA, but less severe.

+Detrended Correspondence Analysis (DCA)

Developed in response to the arch effect in CA. The arch effect is eliminated by detrending using polynomials and segments. Using polynomials is better: a regression is performed in which the second axis is a polynomial function of the first axis and the second axis is replace by the residuals of that regression. This is repeated for any subsequent axes. Sometimes this approach doesn't completely remove the arch effect and detrending by segment is used. In this approach, points along the first axis are split into groups and are centered to have a zero mean on the second axis. This post hoc detrending is desirable as it removes the arch effect, but it isn't the most elegant approach.

Useful sources of information:

-The ordination webpage at Oklahoma State University (http://ordination.okstate.edu); A good "digested" source

-Numerical Ecology

Legendre & Legendre 1998; A huge detailed source that is difficult to digest at times