Multivariate Analysis EBIO 5460
Exercise 4
Eigen-based Ordination (PCA, DCA, CCA)

PRINCIPAL COMPONENT ANALYSIS (PCA)

PCA is an unsupervised method, meaning that it looks to account for the greatest sources of variation regardless of the data structure. PCA is an ideal technique for data that has linear relationships among variables; in other words, how well relationships can be represented by straight lines. Ecological community data are rarely amenable to PCA, with relationships between environmental variables likely to be more linear than relationships among species. Genomic data can often have linear relationships (we will hear more about this next week). The take home point here is that if you use this technique, make sure you justify a linear model. If beta-diversity is low, you might be able to get away with it. A horseshoe or arch of points is a warning sign of poor fit. Beware!

There are many PCA methods available for R. Here we will focus on the most basic one (prcomp), a default function from the R base package.

Let's preform a PCA, again using the dune dataset (I know, it is getting boring, but it is nice to compare with what we have done in the previous exercises and sometimes boring is a good place to start).

```
library(vegan)
data(dune) # loading the dune data from vegan again!

dune.rel <- decostand(dune,"total")
dune.pca <- prcomp (dune.rel, scale=TRUE)
biplot(dune.pca)
```

Before going further, make sure you can distinguish between the species loadings and site scores. What does the default output in R give you? How is each shown in the plot?

PC axes are ordered by decreasing amount of variance explained. So PCA1 explains the most variance in the dataset – this doesn't necessarily mean that that this is the variance that you are most interested in, or is more meaningful to your question, but it oftentimes is. PC axes are orthogonal, meaning they are uncorrelated from one another. The pca summary gives you more information:

```
summary(dune.pca)
```

Make sure you note the proportion of variance explained by each component, you will need to report this in your results section, and it will help you decide how many components to examine.

Missing values are problematic, and are best to deal with prior to the analysis (using approaches we have discussed in class). There is an option in prcomp (na.action) if you haven't dealt with missing values before analysis. The default is na.omit, which removes the whole row with a missing value. Make sure that is something you want to do!

Note that above, scale was set to true. This scales variables to have unit variance before analysis takes place (i.e., divides the centered columns by their variance), which is equivalent to using a *correlation* matrix. Do another PCA but drop scale=TRUE, and you will notice that the range of values within a variable can also affect PCA results (compare location of site 1).

```
dune.pca2 <- prcomp (dune.rel)
biplot(dune.pca2)
```

If you think of PCA as maximizing variance explained, then it makes sense that if some variables show more variance across sites, that might affect results. That is, when we calculate a PCA from a *covariance* matrix, some species contribute more to the total variance. Just make sure that this higher variance isn't something that you want to describe!

Centering is also a default in the analysis. Centering subtracts the mean of the variable, so the mean value becomes 0. One way to think about this is that you are transforming the space to have a new origin. You should almost always do this, as you don't want differences in the mean to affect the loadings (e.g., you don't want variables with a large mean value to matter more than some with a smaller mean).

Of course you could do both of these prior to analysis. Also an important note: PCA is a parametric multivariate analysis (all ordinations we use this week are). You should consider transforming data to try to approximate normality by assessing skewness and kurtosis. Skewness is something in particular to look into with species data, as it is often positive (the right tail is too long). This is a tough assumption to test with so many variables. McCune and Grace (2002) suggest a rule of thumb to try to have the absolute value of skewness to be less than 1. If skewness is too high, try a log transformation.

Outliers are also very influential in these analyses (more so than NMDS), and sometimes a first axis of a PCA is devoted to describing the separating of one outlier from the main cloud of points. This does not look to be the case here.

For alternative methods for computing PCA, please refer to your textbook, Numerical Ecology with R. A vegan tutorial using similar methods as your textbook can be found here: cc.oulu.fi/~jarioksa/opetus/metodi/vegantutor.pdf

DETRENDED CORRESPONDENCE ANALYSIS (DCA)

DCA is often viewed as an improvement on PCA to address issues of linearity. However, it is a brute force method that slices and rescales a Correspondence Analysis (CA: based on Chi-

square distances), and it is therefore hard to interpret. Below, we show how to conduct a DCA so that you can see how it works, but we do not recommend that you use this method:

```
dune.dca<-decorana(dune.rel)
```

To see results, call dune.dca

```
dune.dca
```

To see more results, use

```
summary(dune.dca)
```

DCA is in vegan, and so you can use the same functions to plot as you did with NMDS. The plot function automatically accesses the scores. Make an ordination plot.

```
plot(dune.dca)
```

And go from there based on what you know from last module…. add an ellipse or convex hull? If you want more, function `ordirgl` (requires **`rgl`** package) provides three-dimensional graphics that can be spun around or zoomed into with your mouse. Function `ordiplot3d` (requires package **`scatterplot3d`**) displays simple three-dimensional scatterplots.

CONSTRAINED ORDINATION – CANONICAL CORRESPONDENCE ANALYSIS (CCA)

Constrained ordinations have conditioning terms that are "partialled" out. If you add many constraints (e.g., environmental variables) you move closer to a solution similar to unconstrained ordination. It is best to choose constraints *a priori* based on your knowledge about the biology and your question. Here, we show how to conduct a CCA, which assumes that species (or attributes) have a unimodal relationship with each environmental variable:

```
data(dune.env)
dune.cca<-cca(dune.rel~A1 + Management, data=dune.env)
dune.cca
plot(dune.cca)
```

You can also call a summary (dune.cca) to get more results information. Note that biplot arrows represent the conditioning terms and can be interpreted similarly to other ordinations that involve environmental variables.

In vegan, you can test the significance of the constraints using a permutation test that mimics the standard ANOVA function.

```
anova(dune.cca, by= "term", permutations=999)
```

Note that this test is sequential, the terms are analyzed in order that they are in the model in your call above. So in this case, A1 before Management.

You can also analyze significance of each axis:

```
anova(dune.cca, by= "axis", permutations=999)
```

Remember that this test analyzes variance in general in your dataset, not variance specific to your question.

If you want to account for a variable (treat it as noise), you can also specify a term that are partialled out from the analysis before the constraints. For instance, below you can remove the effect of moisture before analyzing the effects of A1 and Management:

```
dune.cca2 <- cca(dune.rel~A1 + Management + Condition(Moisture), data=dune.env)
dune.cca2
```

This affects the significance of the other terms:

```
anova(dune.cca2, by="term", permutations=999)
```

Above, we calculated CCA and DCA using relativized data so that the ordinations can be directly compared with other types of ordinations that we conducted previously. However, it can also be argued that the transformation used to obtain Chi-square distances for CA-based ordinations is a type of relativization and that it is unnecessary to relativize beforehand. How do the results change when you use the raw data?

Lastly, redundancy analysis (RDA) is analogous to a CCA, only it relates environmental variables to a PCA instead of a CA. This type of constrained ordination should be used if species (or attributes) have linear relationships with environmental variables instead of unimodal relationships. See your textbook or the vegan tutorial referenced above if you want to learn more about this method.

ON YOUR OWN. Explore one of these approaches (PCA, DCA, CCA) with your dataset. Compare your results to what you found with NMDS. Is there a reason to pick one over the other?