**Ordination Methods for Multivariate Data**

This guide will describe how to perform some basic ordination and dimensionality reduction methods in R. This guide is designed to be a living document for Burghardt Lab members and friends. Please edit, comment, expand.

***Overview***

First, we are going to focus on Principal Component Analysis (PCA) and Redundancy Analysis (RDA). Both of these are methods for analyzing gaussian (normally) distributed data. Eventually, Gina will add in an exploration of Non-metric multi-dimensional scaling (NMDS) and Principal Coordinate Analysis (PCoA) which are methods that work with a broad range of data distributions by relying on calculations of rank order (NMDS) or a variety of "distance" metrics (PCoA) including Bray-Curtis Dissimilarity which is frequently used for community composition data. It is important to remember that all of these ordination methods look for VARIATION between samples. Thus, they DO not tell you, for instance, which strains of bacteria are most common overall in your data--just which ones vary the most among your samples (PCA) or Treatments (RDA).

***Why perform ordination in the first place?***

1. It is impossible to visualize more than three dimensions simultaneously. This can make it hard for us to find important patterns in data.
2. The axis defined by the traits and data measured directly may not be the most important to the organism.
3. Running a single multivariate analysis saves time and is often more statistically rigorous than performing lots of separate analysis. For instance, this solves the issue of multiple comparisons.

If you want to read more comprehensively about ordination methods here are a couple of links:

* <https://mb3is.megx.net/gustame/home>
* <http://ordination.okstate.edu/overview.htm>

***PCA Overview:***

A PCA analysis simply reduces the dimensionality of your data by defining new PC axes that more efficiently captures the (co)variation in your dataset than your original traits. Performing a PCA on data gives you a first look at how variation is structured between replicates in a multivariate (normal) data set. This could be a dataset where you have 15 traits measured on each of many plants or it could be a dataset where you have the fitness of 68 strains of bacteria measured in soil mesocosms. PCA is always the first analysis I do and is "unconstrained" and "exploratory" because no covariates are considered.

As Jenn astutely asked, this transformation does not necessarily reduce the overall number of dimensions in your data, but it does concentrate the variation into the top axis so you can focus on the main ways your data varies. If two traits (say for instance height and width) are highly correlated in your data set they will be defined along the same PC axis.

First, read this (much better) explanation of PCA on the Explained Visually site: <https://setosa.io/ev/principal-component-analysis/>

There are multiple ways to perform PCA in R (for discussion see [http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/](http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/))).

I think one of the easiest to work with is the library "FactoMineR" and the helper library "factoextra". I will demonstrate that package in the example below

***RDA overview:***

Redundancy analysis is a method to extract and summarizes the variation in a set of multivariate response variables that can be explained by a set of explanatory variables. The addition of explanatory variables makes it a "constrained" analysis that allows you to estimate the amount of variation in your multivariate outcome that is explained by covariate(s) of interest. RDA has two steps under the hood. First, it runs a multivariate regression to estimate the effect of each explanatory variable on each response variable. Then it performs a PCA on the matrix of fitted values to summarise the dimensions in which the response variable changes in response to the explanatory variable. The RDA (or constrained) axis describes the variation due to the variable of interest. The remaining variation in the data is summarized in addition "unconstrained" PCA axis. If this is clear as mud, take a look at this attempt to describe it ( <https://mb3is.megx.net/gustame/constrained-analyses/rda>) or this one (<http://ordination.okstate.edu/overview.htm#Redundancy_Analysis>).

In the example I will go through our explanatory variable will be one of three temperature treatments and our response variable will be the fitness of 68 strains of Ensifer.

The "vegan" package in R provides many of the functions necessary to perform RDA and I will use those in the analyses below.

As a side note, the direct interpretability of the eigenvalues that emerge from an RDA provides part of the advantage of RDA over types analyses. Eigenvalues are an important concept in linear algebra. If you feel like attempting a Deep Dive to try to understand this better, here is another "Explained Visually" post about Eigen Vectors: <https://setosa.io/ev/eigenvectors-and-eigenvalues/>

***The Tutorial Dataset:***

The data we will use to explore these methods is a strain frequency dataset gathered by Liana Burghardt and Brendan Epstein in Peter Tiffin’s Lab @ UMN. We inoculated 104 sterile soil mesocosms with an initial mixture of 68 strains of *Ensifer meliloti.* This is unpublished data so please do not share beyond this group.

There were eight treatments:

* Field Soil at 22C (F),
* Clay (Cl),
* Drought (Dr),
* salt addition (Na),
* small field soil particles (Sm),
* large field soil particles (La),
* Field Soil 4C (F4),
* Field Soil 32C (F32),

We measured frequencies at three time points—

* 2 weeks (2wk),
* 2.5 months (2.5m),
* 6 months (6m).

Now open the annotated R code file “OrdinationMethodsTutorial\_Mesocosm\_RDA&PCA.R” and work through the annotated code. Note this is a work in progress!

***Questions:***

1. Is temperature an important determinate of strain fitness in our mesocosms? How much variation in your data is explained by the treatments?
2. Which strains are associated with success in 32C? What about 4C?
3. How similar are the results from the “exploratory” PCA and the “explanatory” RDA analysis? In what circumstances would we expect them to be more divergent?

***Additional Questions***

1. In the tutorial, I only visualized the first two major axes of variation. How much variation is explained by the third and fourth axes in the PCA? Can you figure out how to plot them?
2. Which temperature environment is most selective? One way to quantify this is to use the functions range, sd, and skewness to summarise the distribution of strain fitness in each treatment. Can you write some additional code by modifying the summarize function to quantify this? If you want to formally test for statistical differences you can calculate these metrics on every sample and use an ANOVA to test for differences between treatments.
   1. **Range** is perhaps the simplest measure of the strength of selection. It is the difference between the fitness of the best and worst strain. The bigger the difference the stronger selection.
   2. **Standard Deviation** measures the square root of the variance. The larger the variance around the mean the stronger selection.
   3. **Skewness** is the degree to which fitness is asymmetric around the mean (remember in a true Gaussian distribution the mean and the median are the same). Positive (or right) skewness indicates that more strains have fitness lower than the mean then in a true Gaussian distribution (the median is lower than the mean). This means there is selection for a few awesome strains. In contrast if the skew is positive, this means that there are lots of strains with slightly positive fitness and a few strains that basically disappear in the mesocosms entirely.
3. Write some code to identify “generalist” strains that do well in all temperatures and “specialist” strains that do well in only one.
4. How does strain fitness change over time within a treatment?
5. Is Time or Temperature a more important driver of strain fitness in our mesocosms?
6. How do these results differ from the inferences we would make from ordination methods that do not require a normally distributed outcome. What would happen if we used NMDS or PCoA to analyze the raw frequencies?
7. Pick an additional set of treatments (e.g. Drought/No Drought; soil particle size) and analyze those comparisons in the same way that you did Temperature