**Ordination**

Most work on ecological communities involves collecting data at multiple sites on the occurrence and abundance of multiple species, with an objective of determining whether there are any patterns of diversity by site. In addition, multiple environmental variables are often measured at these sites to try to determine what environmental factors account for the abundance and distribution of species. These objectives are more complicated than they appear at first glance because of the large number of non-independent variables. **Ordination** is the term for a family of techniques that summarize multivariate data into fewer dimensions than the original dataset, and then graphically display differences among samples. Ordination is used to quantify the interrelationships among multiple interdependent variables and to explain those variables in terms of a smaller set of underlying dimensions (called components). This involves condensing and information contained in the original variables into a smaller set of dimensions with minimal loss of information. The overall objective is to generate a reduced number of new, synthetic axes of variation that are used to display the distribution of data objects along gradients in your dataset.

To put it another way, biodiversity data are inherently multivariate, with many variables that are correlated with each other. These interrelationships complicate determining which variables are the most important drivers of patterns in your response variables. Moreover, including numerous variables that are essentially redundant to each other can create spurious patterns that would not be present in a more independent set of data. To summarize complex relationships and to identify the most important patterns from a near-infinite number of possible arrangements of multivariate data, the ordination family of techniques can be used. There are *many* types of ordination, each with different purposes, data assumptions, and algorithms, but all result in placement of objects along an axis (an “ordinate”). Ordination literally means arranging items along an axis or axes. Quite often, a few axes account for the majority of the variance in your data; these axes are synthetic combinations of variables. Thus, large, multivariate datasets can be displayed in just a few (often 1-3) dimensions. Ordination allows us to select the most important factors from multiple possibilities, separate strong patterns from weak ones, and reveal unforeseen patterns. Thus, ordination is used to seek and describe patterns in your data.

There are two ways of doing this: **unconstrained** and **constrained** ordination. Constrained ordinations attempt to explain the variation in a set of response variables (e.g. species occurrence or abundance) by the variation in a set of explanatory variables (e.g. environmental variables) measured in the same set of sampling units (e.g. sites). The output includes an ordination plot that only displays the variation in the explanatory variables. The final ordination scores are influenced by the relationship of the response variables to a set of explanatory variables as well as by patterns inherent to the response variables. In contrast, unconstrained ordinations only examine the response variables and display overall variation in the data. An unconstrained ordination is useful for viewing overall patterns in the data, particularly when you do not have a firm idea of which variables are the most important. Constrained ordinations allow you to test hypotheses of how environmental variables determine response variable values, and to discover trends that would be hidden in an unconstrained ordination (i.e., masked by high variability and/or high correlation structure). When you are comparing an unconstrained and constrained ordination on the same data, it is important that you use the same distance metric (explained below) in order to compare them equitably.

**Unconstrained** **ordinations** are carried out on a single data table at a time to detect patterns in the samples and to identify variables responsible for those patterns. (Principal Components Analysis, which we’ll cover next time, is an example of this, as is Nonmetric Multidimensional Scaling, which we’ll cover later, but there are others that we don’t have time to cover, such as Correspondence Analysis [also called Reciprocal Averaging], Detrended Correspondence Analysis, Principal Coordinates Analysis, and Polar Ordination, among others.) They are thus a type of **indirect gradient analysis** whereby interpretation of cause-effect (driver-pattern) can only be made indirectly because those factors were not explicitly included in the analysis.

**Constrained ordinations** are carried out to identify relationships among two or more such datasets simultaneously. This is done by ordinating the first dataset on axes that are combinations of the second dataset; gradients in the first dataset can be described directly in terms of the second, explanatory dataset. A constrained ordination is thus a **direct gradient analysis** because it includes two or more different sets of ecological information into a single analysis, using this additional data to direct (guide) the analysis of the dataset of measured variables. (Recall that constrained ordinations utilize only one dataset of measured variables.) For example, if you collect species abundance and some environmental variables in the same sample (e.g. at the same site), you have two sets of variables that could have a cause-effect relationship because organism abundances are considered a function of environmental conditions. Common constrained ordinations in ecology are Redundancy Analysis, Canonical Correlation, Canonical Discriminant Analysis, and Canonical Correspondence Analysis. All these methods can be lumped under the general term canonical analysis in that they are direct comparisons of more than one data matrix. (“Canonical” is a word with a specific meaning in statistics, referring to techniques meant to find relationships between sets of variables by searching for laten [i.e., hidden] gradients that associate these sets of variables. Many multivariate analyses have the word canonical in them, including Canonical Correlation Analysis, Canonical Correspondence Analysis, and others.)

A primary goal in any type of ordination is to produce a graph that allows you to detect potential patterns in your data. The output of an unconstrained/indirect gradient ordination displays the overall variance in the data whereas the output of a constrained/direct gradient ordination displays only the variation that can be explained by the constraining variables.

**Ordination: the procedure**

To perform most forms of ordination, the similarity/dissimilarity of variables must first be assessed. This is needed because on an ordination plot, objects that are displayed close to each other represent objects that are similar to each other. Thus, we need to first calculate the “distance” between objects. This can be done in many ways: traditional correlation or covariance analysis are two such ways. But there are other ways as well that do not have the same assumptions as correlation/covariance and can be used for other analyses.

**Similarity, dissimilarity, and distance:**

**Similarity** is a characterization of the ratio of the number of attributes two objects share in common compared to the total list of attributes between them. Objects that have everything in common are identical, with similarity = 1. Objects with nothing in common have similarity = 0.

**Dissimilarity** is the complement of similarity and is a characterization of the number of attributes two objects have uniquely compared to the total list of attributes between them. In general, dissimilarity can be calculated as 1 – similarity.

**Distance** is a conception of the “proximity” of objects in a high dimensional space defined by measurements on the attributes. **This is not the same thing as how the word distance is normally used to indicate spatial proximity** (e.g. the distance between Lubbock and Dallas is ~530 km). Instead, it quantifies the **dissimilarity** of portions of your dataset(s): objects that are similar have small distances between them whereas objects that are different have large distances.

In practice, these terms are sometimes used interchangeably. They have distinct properties, however. Dissimilarities are bounded [0,1]; once plots have no species in common they can be no more dissimilar. The same goes for similarities (of course). Distances, on the other hand, are unbounded on the upper end; plots that have no species in common have distances that depend on the number and abundance of species in the plots, and is thus variable. I tend to use just the term distance to serve as a blanket term covering distance, dissimilarity, and thus, by extension, similarity.

There are many ways of calculating distances; I will cover the different types below. R calculates distances with the dist()function, which only provides a few choices. The *vegan* package uses vegdist()and *labdsv* uses dsvdis(), which together provide a larger number of possible distance metrics. The table below lists them; X indicates the distance metric is present under that name whereas an entry other than X indicates that the distance is present in the function but is called something slightly different; consult R or the package’s documentation/help for specific terminology.

(And there are also several other R packages that can be used to calculate distance.)

|  |  |  |  |
| --- | --- | --- | --- |
| **Distance Metric** | **R** dist() | ***vegan*** vegdist() | ***labdsv*** dsvdis() |
| Euclidean | X (default) | X |  |
| Manhattan | X | X |  |
| Mahalanobis |  | X |  |
| binary | X |  | steinhaus1 |
| Sørensen |  |  | X1 |
| Canberra |  | X | X |
| Bray-Curtis |  | bray (default) | bray/curtis |
| Gower |  | X |  |
| Kulczynski |  | X |  |
| Ochiai |  |  | X |
| Ruzicka |  |  | X |
| Roberts |  |  | X |
| Chi-Square |  |  | chisq |
| Morisita |  | X |  |
| Mountford |  | X |  |
| Horn |  | X2 |  |
| Minkowski | X |  |  |
| Etc. |  | X |  |

1 = converts data to presence/absence

2 = Horn-Morisita Index, not the Horn Index (which has a different formula)

(Some of these measures can also be used to assess beta diversity; you may wish to refresh yourself about this topic from the “Diversity indices” lesson from a few weeks ago.)

Some types of ordination require a specific metric; for example, PCA uses Euclidean, Correspondence Analysis uses Chi-square, etc. But for most forms of ordination, you have to pick a metric. The type of distance metric you use can have a profound effect on your results. Thus, an obvious question springs to mind: **What distance metric should you use?**

In most situations, there may not be a clear-cut answer to this question. There are dozens of distance metrics, and for any given objective, there may be many that will yield similar and accurate answers. You should first examine the structure of your data (does it contain both categorical and continuous variables? does it contain lots of 0’s? etc.), and then the information below can help you decide. Note that although there may not be a right answer to this question (many may be appropriate for your dataset), there may be some wrong answers (metrics that are not appropriate, given your data).

(If you want details on the math behind each of these, type ?dist and ?vegdist in R.)

Some (but by no means all) examples:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Gower coefficient** | | | | | Useful for heterogeneous data sets (i.e., mixed categorical and continuous data). It calculates a partial similarity value of two objects for each variable describing them. The final similarity score is the average of all partial similarities. Binary, qualitative and semi-quantitative, and quantitative variables are treated differently. Handles missing data. | | | |
| **Steinhaus coefficient** | | | | | Widely used forraw count data. It compares the sum of the minimum, per-variable values between two objects to the average value of all variables describing these objects. If applied to binary data, this is equivalent to the Sørensen coefficient. | | | |
| **Euclidean distance** | | | | Not good in gradient separation unless data have been standardized (e.g. via decostand() in *vegan*, which you did in the “Site x environment” lesson to divide by row totals; type ?decostand for more options of other types of standardization it can do). The more variables present in a data set, the larger that Euclidean distances will be. Appropriate for when you have continuous numerical variables and want to reflect absolute distances. It works even if variables are correlated and if different scales of units are used for different variables. However, it often does not work well with diversity data because very high beta diversities (turnover) generate inflated distances (Mahalanobis distance is generally recommended in such cases). Moreover, it does not handle 0’s well (works best if <20% of cells in your dataset are 0’s), although it will run even if you have empty rows or columns. Sensitive to outliers. Used in PCA. | |  | | | |
| **Mahalanobis distance** | | | | Eliminates the effect of correlations between variables and is arrived at through the calculation of a covariance matrix from the input matrix. It also eliminates differences in scale between variables. | |  | | | |
| **Manhattan metric** | | Similar to Euclidean distance but rather than using the Pythagorean theorem, the Manhattan distance simply sums the absolute differences across pairs of variable values for a given object. Just like the Euclidean distance, this metric doesn’t handle 0’s well and is affected by sample size. And it too is not good in gradient separation unless data have been standardized (e.g. via decostand() in *vegan*, which you did in the “Site x environment” lesson to divide by row totals; type ?decostand for more options of other types of standardization it can do). Handles high beta diversity better than Euclidean. | | | |
| **Canberra metric** | | This metric excludes double zeros and increases the effect of differences between variables with low values or many zeroes. | | | |
| **χ2 distance** | | The Chi-square distance is used in various forms of correspondence analysis (like Canonical Correspondence Analysis that we will be doing later in the semester as well as Correspondence Analysis and Detrended Correspondence Analysis). The Chi-square distance reduces the influence of very high values in a dataset, meaning that it de-emphasizes common species. Cannot be used if there are negative values in your dataset. It will run even if you have empty rows or columns. | | | | | |
| **Hellinger distance** | | Appropriate when you want to emphasize differences between variables. Can be useful when you have very different sample sizes and lots of 0’s in your dataset. Variables with few non-zero counts (such as rare species) are given lower weights. Often used for abundance data. | | | | | |
| **Bray-Curtis dissimilarity** | | Appropriate when you want to differentiate sampling units and also take relative magnitudes of data into account. It likewise handles 0’s well. This measure treats differences between high and low variable values equally. Often used for NMDS. This is the default metric used in *vegan* because it is one of the most flexible metrics. | | | | | |

There are many others; the ones I’ve included above are the most commonly used ones.

**Now that I’ve chosen which distance index I will use, how do I create a distance matrix?**

Open a new RStudio session (with your class working directory) with the following libraries:

*labdsv*

*MASS*

*MVA*

*optpart*

*picante*

*stats*

*vegan*

Read in bryceveg.R:

veg <- read.table("bryceveg.R", header=TRUE)

This is the syntax used in different packages to create a distance matrix from veg, using various distance metrics:

#In R:

result <- dist(veg, method="canberra")

#is the same as:

result <- dist(veg, "canberra")

#In *vegan*:

result <- vegdist(veg, method="bray")

#is the same as:

result <- vegdist(veg, "bray")

#In *labdsv*:

result <- dsvdis(veg, index="ruzicka")

#is the same as:

result <- dsvdis(veg, "ruzicka")

One convention is to name the result "dis" with an extension that specifies which distance index you used. For example (for R, *vegan*, and *labdsv*, respectively):

dis.eu <- dist(veg, "euclidean")

dis.bc <- vegdist(veg, method="bray")

dis.ruz <- dsvdis(veg, index="ruzicka")

If you have binary data, you should specify:

dis.bc.bin <- vegdist(veg, method="bray", binary=TRUE)

You may also wish to include something more descriptive about the data in the name as well:

bryceveg.dis.bray <- vegdist(veg, method="bray", binary=TRUE)

Use the Bray-Curtis distance index on bryceveg.R to create an R object called dis.bc:

dis.bc <- vegdist(veg, method="bray")

**Now that I have a distance matrix, what type of ordination should I use?**

The choice of which of these you should use depends on: 1) the type of data you have, 2) the distance matrix you want/can use, and 3) what your objectives are. At this point, I recommend that you read Paliy and Shankar (2016) (<http://www.wright.edu/~oleg.paliy/Papers/Paliy_ME2016.pdf>) about which method to choose. Although their paper is on microbial communities, it applies more broadly to any ecological community.

All of these methods have the same broad goal; they differ in their algorithms/approaches and, thus, in how they are influenced by the structure of your data.

*Indirect Gradient Analysis/Unconstrained Ordination -* Gradients are unknown *a priori* (usually are inferred from species composition data). Good for when you do not have site x environment data, only site x species. Examples include:

* PCA: Principal Components Analysis (which we’ll cover next time) – based on a Euclidean dissimilarity between samples; subject to **arch effect** (distortion in ordination plot caused by unimodal [rather than linear] distribution of species along gradients) and **horseshoe effect** (distortion in ordination plot if environmental gradient is too long); prcomp() and princomp() in *stats*, pca() in *labdsv*; princomp() only works if you have more rows (sites) than columns (variables)
* NMDS: Nonmetric Multidimensional Scaling (which we’ll cover in a few weeks) – uses ranks to carry out the correlation between distances in a dissimilarity matrix and the distances in a low-dimension space, which reduces arch effect; can be used with presence/absence data; works well if you have species with very large differences in abundances and is not affected by outliers, clustering in your data, or a moderate level of noise; however, it does not find unique solutions, you must designate dimensions *a priori*, and its algorithm can become trapped in a local minimum; isoMDS() in *MASS*, monoMDS() in *vegan*, nmds() in *labdsv*
* CA: Correspondence Analysis (also called RA: Reciprocal Averaging) – is similar to PCA and likewise subject to arch effect but handles long ecological gradients better than does PCA; cca()and decorana() in *vegan*
* DCA: Detrended Correspondence Analysis – removes arch effect from CA; can handle large, complex datasets along long ecological gradients as well as nonlinear and unimodal data (i.e., handles data where PCA fails); axes of ordination plots are in units of beta diversity; is sensitive to outliers and performs poorly with skewed data; decorana() in *vegan*
* PCO or PCoA: Principal Coordinates Analysis (also called MDS: Multidimensional Scaling) – maximizes linear correlation between distances in a dissimilarity matrix and the distances in a low-dimension space; if you use a Euclidean distance matrix, the results are the same as PCA; subject to arch effect; can use presence/absence data; cmdscale() in *stats*, pcoa() in *ape*
* PO: Polar Ordination – used to examine the organization and relationship among entities on a pre-determined and well-defined gradient without requiring multivariate normality; however, this method is seldom used anymore because it is highly subjective in assigning endpoints (poles) in the gradient, it is subject to severe distortions and erroneous conclusions if outliers are designated as endpoints, and is subject to an arch effect; polar.ord() in *asbio*

*Direct Gradient Analysis/Constrained Ordination*- Any gradient analysis in which the important gradients are known *a priori*. Appropriate for when you have both site x species and site x environment data. Examples include:

* RDA or RA: Redundancy Analysis – similar to CCA (below) in that it uses environmental variables as explanatory factors, but it uses absolute rather than relative abundances, and it assumes monotonic (usually linear) rather than unimodal responses; rda() in *vegan*
* CanCor: Canonical Correlation – best used only as an exploratory method to make sense out of an otherwise unwieldy number of bivariate correlations between sets of variables; has several limiting assumptions (linearity, normality, absence of multicollinearity); seldom used now; cancor() in *stats*
* CCA: Canonical Correspondence Analysis (which we’ll cover in a few weeks) – uses environmental variables as explanatory factors; maximizes the correlation between species scores and sample scores, but the sample scores are constrained to be linear combinations of the explanatory variables; uses relative rather than absolute abundance, so is not sensitive to rarity; constrained form of RA; handles noisy and skewed data better than DCA (although if data are skewed, it may be preferable to do a square root or log transformation to prevent a few high abundance values from dominating a variable’s contribution); is not hampered by multicollinearity or high correlations between X or Y variables; cca() in *vegan*

**General recommendations for ordination: A dichotomous key**

(from Michael Palmer, Oklahoma State University)

You first have to determine whether you want/need to perform a direct (constrained) or an indirect (unconstrained) gradient analysis: that’s couplet 1. Then based on your answer, follow the subsequent couplets:

1. Direct Gradient Analysis.................................................2

2. Few species.....................................................................4

4. Monotonic responses to gradients................................Linear regression

4. Nonmonotonic responses to gradients.........................Generalized linear models

2. Many species..................................................................5

5. Monotonic responses ..................................................RDA (Redundancy Analysis)

5. Nonmonotonic responses.............................................6

6. concerned about arch effect........................................DCCA (Detrended Canonical

Correspondence Analysis)

6. not concerned about arch effect..................................CCA (Canonical Correspondence

Analysis)

(An arch effect is when the second axis of an ordination is an arched function of the first axis. It is caused by the unimodal distribution of species along gradients. The first axis is composed of the most important variables, the 2nd axis by the next most important set, etc.)

1. Indirect Gradient Analysis...............................................3

3. Only distance values are available..................................7

7. Monotonic responses .................................................PCO (Principal Coordinates

Analysis)

7. Nonmonotonic responses...........................................NMDS (Nonmetric Multidimensional

Scaling)

3. Raw data are available...................................................8

8. Monotonic responses ..................................................9

9. Variables noncommensurate…………………..........PCA (Principal Components Analysis)

using a correlation matrix

9. Variables commensurate............................................PCA using a covariance matrix

8. Nonmonotonic responses............................................10

10. Feel OK about prespecifying number of dimensions,

not worried about local optima, not interested in

species scores...............................................................NMDS (Nonmetric Multidimensional

Scaling)

10. Not as above, but willing to accept either arch

effect or detrending/rescaling......................................11

11. Don't like arch but detrending OK ........................DCA (Detrended Correspondence

Analysis)

11. Arch OK, or only interested in axis 1………........CA (Correspondence Analysis)

If a method isn’t included in this key (e.g. Polar Ordination), then it is not generally recommended.

Every ordination method uses a distance matrix constructed on your data, using different methods (e.g. Euclidean, Bray-Curtis, Jaccard, Manhattan, Mahalanobis, etc.) to calculate the distance between samples. These different methods of calculating a distance matrix give different results. Different ordination methods use different distance matrices. For example, PCA uses only Euclidean distance, whereas NMDS and PCO use any distance. (Methods that use the same type of distance matrix should yield the same results, e.g. PCA and an NMDS built with a Euclidean matrix.)

So here are some rules of thumb about which method to use:

- If you have only site x species data, you are restricted to using an unconstrained form of ordination; if you have site x species and site x environment data, then you can use constrained forms of ordination.

- If your objective is to provide a quantitative description of the major ecological gradients of variation among individual sampling entities and/or to portray sampling entities along continuous gradients of maximum sample variation, then PCA, PCO, RDA, CA, DCA, NMDS, and others not mentioned here can be used.

- If you determine (or assume) that there are linear gradients of variation, then you can use PCA, PCO, RDA.

- If you determine/assume there are unimodal gradients, then CA, DCA are warranted.

- If you determine/assume there is no linear or unimodal relationship, only a monotonic relationship between input and output dissimilarities, then use NMDS.

- If your objective is to provide a quantitative description of the major ecological patterns in a set of response variables explainable by a set of explanatory variables, then CCA and other constrained ordination techniques not mentioned here can be used.

- If you determine/assume a unimodal response function of response variables (species) along gradients defined by the explanatory variables (environment), then use CCA.

- If you have a dataset that includes lots of zeros, then use of a Bray-Curtis matrix and NMDS ordination is recommended. Bray-Curtis distance is chosen because it is not affected by the number of null values (0’s) between samples like Euclidean distance.

- If you have a dataset that does not include a lot of 0’s, then you can use Euclidean distance and use either PCA or NMDS.

We will examine several types of ordination in the coming weeks: Principal Components Analysis, Redundancy Analysis, Non-metric Multidimensional Scaling, and Canonical Correspondence Analysis. Two of these (PCA, NMDS) are unconstrained whereas the other two (RDA, CCA) are constrained forms of ordination. By doing two of each type of ordination, you will be able to see some differences in data assumptions and in output. PCA is typically used to reduce the distribution of a multivariate set of data down to only 2 or 3 dimensions and plot the results, allowing you to identify patterns that can then be modeled linearly. NMDS is useful when you cannot assume a linear relationship. RDA is used to find patterns among the distribution of points in ordination space (e.g. from a PCA) and explore possible environmental factors associated with those patterns. CCA similarly seeks to determine patterns among species-site-environmental variables axes. They differ in that RDA assumes a linear response of species to environmental data whereas CCA assumes a unimodal one. In addition, CCA focuses on the community as a whole (trends in patterns of species composition) whereas RDA focuses more in terms of abundances of species.

All forms of ordination take a large multivariate dataset of correlated variables and reduce it to a set of orthogonal axes composed of independent variables; in doing so, this simplifies and hopefully clarifies relationships in your data by concentrating focus on the most influential variables.

**Ordination as data reduction:**

All forms of ordination are means of data reduction, reducing a dataset to a smaller number of composite (synthetic) variables along continuous axes. These axes are combinations of variables. If you think of a data matrix of *n* entities x *p* variables, that can be reduced to represent the information in a smaller number of dimensions, *k*. Each of the new dimensions is a synthetic variable representing as much of the original information as possible. The only way this can be possible is if the underlying variables covary. Information that is omitted in the reduced dimensional space is the residual variation that represents noise and the influence of minor factors. To quote Gnanadesikan and Wilk (1969):

“The issue in reduction of dimensionality in analyzing multiresponse situations is between attainment of simplicity for understanding, visualization and interpretation, on the one hand, and retention of sufficient detail for adequate representation, on the other hand. Reduction of dimensionality can lead to parsimony of description or of measurement or of both.”

The *k* axes represent the strongest correlation

*n* entities x *p* variables

*n* x *k*

structure in the data. These axes are also

called “principal components.”

After extracting the synthetic variables (axes), one can then examine them in relation to other variables. In community ecology, one usually first summarizes a matrix of sample units (sites) by species into a few axes representing the primary gradients in species composition. Those axes can then be related to measured environmental variables. Ordination thus essentially describes the strongest patterns in species composition, with the underlying independent variables thought to vary continuously along a gradient. (For example, think of how species composition changes along an elevational gradient, with many independent variables such as temperature and precipitation also varying along that gradient.) In **direct gradient analysis**, sample units (sites) are ordinated according to measurements of environmental factors in those sample units. In contrast, **indirect gradient analysis** positions sample units according to covariation and association among species.

**Number of axes:**

How many axes one can extract depends on how many discrete signals can be detected against a background of noise. As the complexity of a dataset grows, so does the number of signals…but so does the noise. Occasionally, you may find a single overriding pattern. More typically, however, you are faced with several possible underlying factors, but their relative importance is unknown, necessitating use of two- or three-dimensional ordinations (with “dimensions” meaning axes). Going beyond three axes is difficult to display and interpret. Each method of ordination has its own way of determining statistical significance of ordination axes.

**Relating variables to ordination results:**

Relating variables to your ordinations (to interpret them), you can use correlation analysis and/or graphically.

-Correlation analysis: you can examine the correlations between axis scores and the other variables (typically, the environmental variables). If a variable has a linear relationship with an ordination axis, it will be indicated in the correlation coefficient, *r*. Values of *r2* (called the coefficient of determination) indicate the proportion of variation in position on an ordination axis that is explained by the variable being examined. (An alternative to Pearson’s *r* is to examine ranked relationships between ordination scores and individual variables by Kendall’s tau.) Correlation analysis is not appropriate for binary data nor when relationships are nonlinear (e.g. a unimodal distribution along an ordination axis). There are other methods that are applicable in those cases, or for examining relationships among multiple variables; see McCune and Grace (2002) for more info.

-Graphical: overlays are good ways of detecting patterns, including non-linear ones. Categorical variables can be displayed with different symbols, and continuous variables can be shown in symbols of different sizes (with size proportional to variable value). See McCune and Grace (2002) for various examples.

**Examples:**

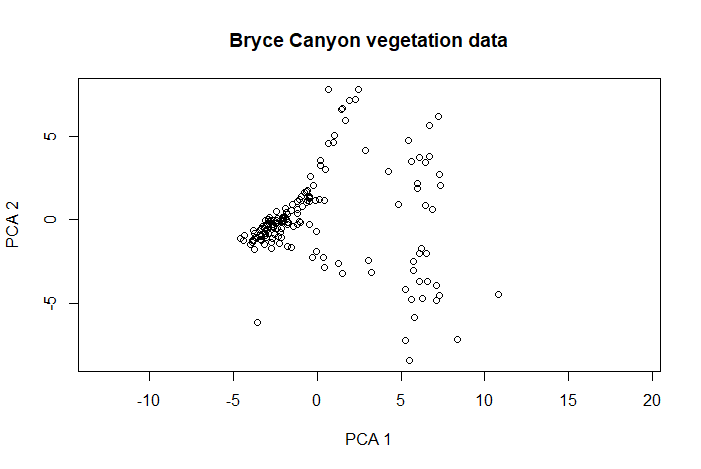
Let’s take the Bryce Canyon vegetation data for a spin, trying out each of the four ordinations we will cover in greater detail in future lessons (we will not go into any depth today, this is just an intro).

PCA:

PCA uses Euclidean distance (default); for this example, the distances are calculated on the data’s correlation matrix:

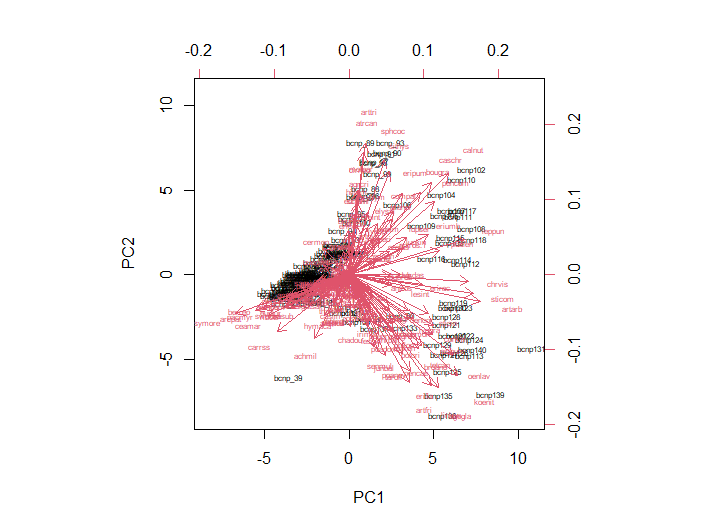
pca.1 <- pca(veg, cor=TRUE, dim=10)

plot(pca.1, title="Bryce Canyon vegetation data")



The resulting ordination plot shows how the sites (circles) fall out on the first two axes, which collectively explain most of the variation in the data. You can see that there is some separation of sites along the first axis, less so on the second. Now let’s examine how species ordinate by sites:

biplot(pca.1$scores, pca.1$loadings, cex=0.5)



This graph is called a **biplot** because it shows two sets of variables (sites and species in this case). The sites are in black text and the species abbreviations are in pink. (The site names occupy the same locations in graph space that the circles did in the previous graph.) The arrows indicate the relationships between species and sites. This graph is kind of a mess because there were 160 sites and 169 species, but you can see (based on arrow placement and length) that some species are more clearly associated with certain sites than are others.

RDA:

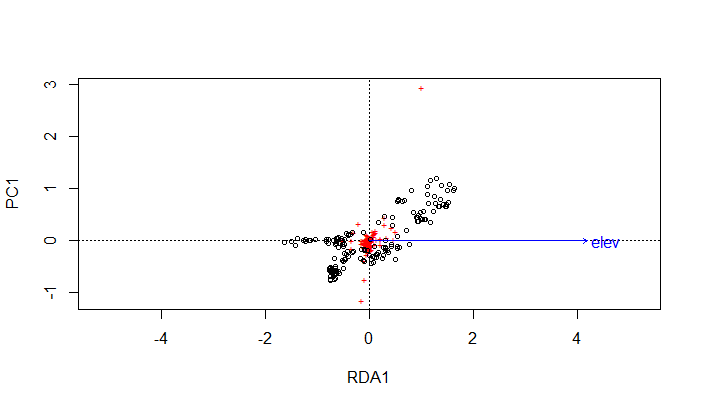
Because this is a constrained ordination, you must first load the Bryce Canyon site x environment data:

site <- read.table("Brycesite.R", header=TRUE)

Then we can examine how an environmental variable (in this example, elevation) can possibly explain patterns in your site x species data:

RDAdemo <- rda(veg ~ elev, site)

plot(RDAdemo)



RDA is a constrained version of PCA. RDA plots an ordination defined by the matrix of response variables and the matrix of explanatory variables. In the plot, the circles are sites, the red crosses are species, and the blue arrow is the environmental variable we are examining (elevation). There is a species that appears to be an outlier (upper center). There is a positive association between the constrained axis (called RDA1) and the first residual axis (PCA1) but not strongly associated with elevation, indicating that there are variables other than elevation that account for patterns in your community.

NMDS:

NMDS works on dissimilarity/distance data, so use the distance matrix you made earlier with the Bray-Curtis metric:

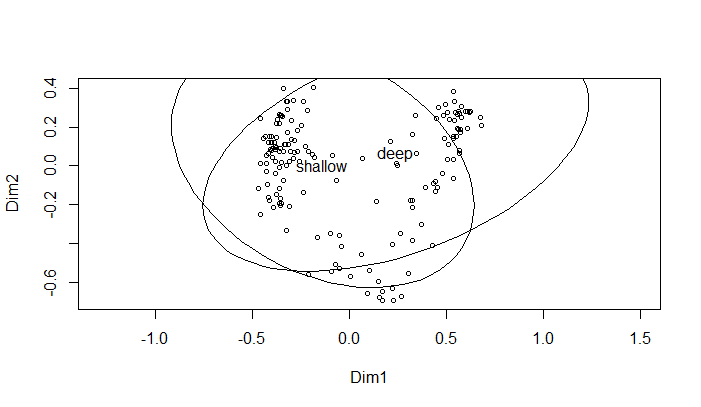
dis.bc <- vegdist(veg, method="bray")

NMDSdemo <- nmds(dis.bc, k=4)

NMDS is a form of constrained ordination, so let’s use it to see whether the site x species data can be separated on the basis of soil depth (since you already have the site data loaded from brycesite.R):

ordiplot(NMDSdemo, display = "sites", type = "points")

ordiellipse(NMDSdemo, site$depth, conf = 0.95, label = TRUE)



The circles are sites separated on the basis of species composition; the ellipses are confidence ellipses. There is considerable overlap in the ellipses, indicating that soil depth is not a factor that accounts for differences in plants by sites.

CCA:

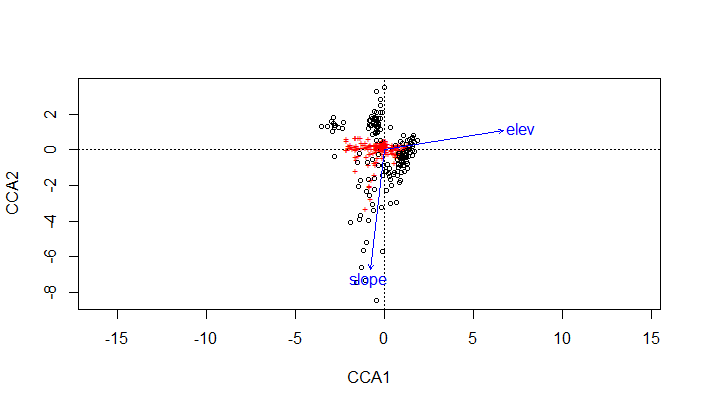
Suppose we want to examine how elevation and slope structure plant species composition in Bryce Canyon:

attach(site)

(The attach() argument allows you to work with columns from an R object, in this case site, without needing to refer to site every time.)

cca.1 <- cca(veg~elev+slope)

cca1.plot <- plot(cca.1)



As with the RDA plot, the species are shown as red crosses, sites are circles, and the environmental variables are blue arrows. The angles between arrows represent associations (correlations) between environmental variables, so this plot shows there is no strong correlation between elevation and slope (we wouldn’t expect there to be one). The direction of an arrow indicates the direction of maximum change in each variable (visually indicating what the axes represent; the cosine of the angle between an arrow and an axis is the correlation coefficient between that variable and that axis). Arrow length indicates importance of an environmental variable. The positions of site points relative to the arrows indicates the environmental conditions at each site; the locations of species points relative to the arrows indicates characteristics of the ecological optima of each species. In this case, it looks like axis 1 is associated most with increasing elevation and axis 2 with decreasing slope. There appears to be more separation of sites by slope than by elevation because there is more vertical separation than horizontal separation.

**References**

Gnanadesikan, R., and M.B. Wilk. 1969. Data analytic methods in multivariate statistical analysis. Pp. 593-638 in: *Multivariate Analysis II: Proceedings of the 2nd International Symposium on Multivariate Analysis, Wright State University, Dayton, Ohio, June 17-22, 1968* (P.R. Krishnaiah, ed.). Academic Press, New York, NY.

McCune, B., and J.B. Grace. 2002. *Analysis of Ecological Communities*. MJM, Gleneden Beach, OR.

Paliy, O, and V. Shankar. 2016. Application of multivariate statistical techniques in microbial ecology. Molecular Ecology 25:1032-1057.

**Assignment:** due 0800 Monday, March 29

Start a fresh RStudio session. Remember to set your working directory to your course folder and use the same package libraries as we used today.

**Q1. Read in Ground\_beetles\_abundance.csv (header = TRUE, row.names = 1). Examine the data and decide which distance metric is most appropriate and why. Using that metric, construct a distance matrix.**

**Q2. Perform a PCA, RDA, NMDS, and CCA on those data, producing plots like the ones produced in the lesson’s exercises for today, with the following changes:**

**For PCA: do not include dim=10**

**For RDA: the site x environment data are in GBsite.csv (header = TRUE, row.names = 1); use that file to examine how beetle abundance is associated with habitat type**

**For NMDS: use k=5 and use GBsite.csv to examine how beetle abundance is associated with habitat type**

**For CCA: use GBsite.csv to examine how beetle abundance is associated with habitat type**

**Q3. Are the results of the analyses you just performed consistent with each other? (In other words, would you draw the same conclusions about ground beetle communities?) If not, how do they differ?**

Make an RMarkdown Word file of your work and turn that in. Be sure to include your answers to any questions asked! Turn in your assignment as a Word document via email to [iroro.tanshi@ttu.edu](mailto:iroro.tanshi@ttu.edu) no later than 8:00 a.m. on Monday of next week. In your email, please include the following as the Subject line:

Assignment on ordination