**Canonical Correspondence Analysis**

Recall that there are many types of ordinations, and the choice of which to use should be based on your objectives and on the data structure. The NMDS we did last time was a form of unconstrained ordination, which is carried out on a single dataset with the objective of identifying patterns in your data and the variables responsible for those patterns. We then used the product in a second analysis with a second dataset to examine the influence of environmental variables on the original ordination.

Other approaches that use environmental variables in the construction of the ordination (thus the site x species dataset AND the site x environmental variable dataset) simultaneously (rather than in a two-step process) are called constrained ordinations because the positions of the samples in the ordination are constrained by the environmental variables. Like other constrained approaches (e.g. Redundancy Analysis, etc.), you thus need two datasets: 1) a dataset of response (dependent) variables (e.g. species presence or abundance), and 2) a dataset of predictor (independent) variables (e.g. environmental variables) measured at the same sites as in dataset 1.

To examine the effects of dataset 2 on dataset 1, one approach is one of the most widely used analytical methods in community ecology: **canonical correspondence analysis (CCA)**. This should not be confused with correspondence analysis, which is a different (and unconstrained) multivariate technique.

CCA allows you to order your sampling units and your species based on relationships between unimodal patterns of redundant co-occurrences in your species data and explanatory variables of your choosing. It is powerful and commonly used but is complex and limited by assumptions that can be difficult to meet in ecological community datasets.

In CCA, you (1) start with a Chi-square distance matrix; (2) regress the differences from expectation on environmental variables to get fitted values, using a weighted regression where total abundance by sites is used as the weights; and (3) calculate the Euclidean distance of the fitted data matrix and project it by eigenanalysis. (Review your notes from the PCA/RDA lesson to refresh your memory on eigenanalysis, eigenvalues, and eigenvectors.)

The importance of specific environmental variables is then assessed by their correlation to the projected scatter diagram. For each sample, except for the data on species composition and abundance, you must have data on environmental factors. As output, for each environmental factor you get its contribution to the explanation of the site x species data distribution and the level of statistical significance. CCA can thus be thought of as a form of multiple regression of species on environmental variables.

**Assumptions of CCA:**

CCA assumes that your data exhibit multivariate normality and homogeneity of variance, and that there is a linear relationship between the canonical variates and each set of variables. Meeting these assumptions typically requires a large sample size. In other words, CCA assumes a unimodal (bell-shaped) response between environmental variables and species composition, which usually is the case when you have large enough environmental gradients so that you have samples from both suitable and unsuitable environments for your species. (If instead you have short environmental gradients and can assume a linear response between environmental variables and species composition, then Redundancy Analysis [RDA] is typically used.)

Interpreting the resulting species scores as representing species "optima" assumes that the abundance of a species is a symmetrical unimodal function of position along environmental gradients. But even when this assumption is violated (the distributions of species along environmental gradients is skewed rather than unimodal), CCA still performs well (see Palmer 1993 for an example). That means that CCA is robust to skewed species distributions and high levels of variability in the data. (However, if the dependent variables are highly skewed, it might be preferable to transform them via a square root or logarithm transformation; this can prevent a few large data values from dominating a variable’s contribution.) CCA also performs well when there is multicollinearity or high correlations between dependent or independent variables. However, it tends to overemphasize the contribution of rare species and underemphasize abundant ones. See McGarigal et al. (2000) for more information.

CCA is a very commonly used multivariate technique in ecology. In contrast to PCA, it uses a Chi-square distance matrix rather than correlation/covariance (or Euclidean distance). This solves the problem of treating joint absences as a positive correlation, and this is an improvement over PCA. On the other hand, it has the problem of spurious "information" when two sites have no species in common, or where their distance is a function of their productivity rather than some constant. Finally, like PCA, it assumes that the modes of species relate linearly to the environmental variables employed in the analysis.

CCA (like every other method covered in this class) is very convoluted, with many variations, applications, and criticisms. You can really go down a rabbit-hole in examining any of them.

In R, there is more than one algorithm to compute CCA. One of these is the cca() function in the *vegan* package.

**Doing a CCA in R:**

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

*labdsv*

*MASS*

*MVA*

*optpart*

*stats*

*vegan*

We will once again be working with the Bryce Canyon site datasets (bryceveg.R and brycesite.R); load them as objects named veg and site, respectively (with header = TRUE for both) and examine the environmental variables in brycesite.R.

Examine site and select those environmental variables you have reason to believe are important and thus will be analyzed as predictors (constraining variables). (In the example below, I have selected elevation, slope, and aspect value as environmental variables that I believe to drive plant community composition in Bryce Canyon.) The choice of which environmental variables you choose to include/exclude will greatly influence the outcome of CCA (and other constrained ordinations). Including all variables can make the process run very slowly, and if you have variables that are correlated with each other, then including all of them will bias your outcome due to overfitting. You should include variables that you think are the most important independent determinants of species composition. But it’s also sometimes desirable to include other variables to explore for unexpected patterns; you can always remove superfluous variables later if they are confusing or difficult to interpret. However, if you’re doing a hypothesis-testing analysis then post-hoc removal of variables to make your results stronger is not valid! If you have at least as many variables as you have samples in CCA, then your ordination is no longer constrained and you will have explained 100% of the variation in species composition, but due to overfitting. In such cases, you should first perform a PCA to identify the most important, independent variables and then use those axes in your CCA (like we did in the PCA lesson where the axes were then used in a constrained ordination, RDA).

attach(site)

The attach() function allows you to work with a dataset without needing to reference it, you just indicate what columns in that dataset you want. However, sometimes you get a warning message: The following objects are masked from [your R object name]. In such a case, use the R object and $ to call out columns (e.g. site$elev + site$slope + site$av).

The site x species matrix goes on the left-hand side of the equation, with your selected environmental variables on the right.

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

#If you had not used attach(site), you would have needed to #express this equation as

#cca(veg ~ elev + slope + av, data = site)

#or as

#cca(veg ~ site$elev + site$slope + site$av)

cca.1

Call: cca(formula = bryceveg ~ elev + slope + av)

Inertia Proportion Rank

Total 10.8656 1.0000

Constrained 0.6976 0.0642 3

Unconstrained 10.1680 0.9358 147

Inertia is mean squared contingency coefficient

Eigenvalues for constrained axes:

CCA1 CCA2 CCA3

0.5194 0.1110 0.0671

Eigenvalues for unconstrained axes:

CA1 CA2 CA3 CA4 CA5 CA6 CA7 CA8

0.6912 0.5571 0.5278 0.4311 0.3752 0.3487 0.2900 0.2605

(Showed only 8 of all 147 unconstrained eigenvalues)

The first items after the Call: give the mean squared contingency coefficients of the analysis. (Because the data matrix is first converted to Chi-square distances in CCA, the entries are contingency coefficients.) The top row of the mean squared contingency coefficients (Total) is the total variability before the matrix is subjected to weighted regression. This is the variability that could be explained. The second line (Constrained) is the variability in the data matrix after weighted regression; this is the variability that is explained by the axes in the CCA. The last line (Unconstrained) is the variance of the residuals of the regression (i.e., unexplained variance).

In this particular example, the CCA was not very informative. Only 0.6975 / 10.8656 or 0.0642 (6.42%) of the total variability was captured in the CCA. Clearly, this analysis was not very successful at capturing the variability in vegetation composition in Bryce Canyon. But again, for pedagogical purposes, we’ll continue with it here.

The next set of lines gives the eigenvalues associated with the projection. The top line gives the constrained eigenvalues. Because we only had three environmental variables (elevation, slope, aspect value), we can only have three constrained eigenvalues. The three values sum to 0.69755, so you can ask R to perform the division for you to determine what proportion of constrained variability is explained:

#CCA1:

0.5194/0.6976

[1] 0.7445528

#CCA2:

0.1110/0.6976

[1] 0.159117

#CCA3:

0.0671/0.6976

[1] 0.09618693

The results show that the first eigenvalue accounts for approximately 74% of the constrained variability, with the second at ~16%, and the third at ~10%. Remember, these are fractions of the constrained variability, which is a small fraction of the total variability in this example. (We’ll touch more on this shortly when I discuss **inertia**.) The first unconstrained eigenvalue is almost as large as the first three constrained values combined here. So these results indicate that the CCA isn’t doing a good job. But we’ll stick with it for illustrative purposes.

The scores() command can get you the scores for samples ("sites"), species ("species"), or constraining variables (called "bp" in R), allowing you to determine how each is represented in each of the first two CCA axes:

scores(cca.1, display="sites")

scores(cca.1, display="species")

scores(cca.1, display="bp")

The scores can allow you to determine which samples are most strongly represented in each axis, and in what manner (positive or negative).

**Inertia** is variance. For linear methods, the inertia represents the variance in species abundances, but in unimodal methods like CCA, it represents the variance or spread of species scores (which you just examined). In the output above, the overall inertia, or variance in species dispersion in the data set, is 10.8656. If there are no covariates, this is equal to the sum of all unconstrained eigenvalues. This is the same value we would have gotten if we had performed a correspondence analysis (an unconstrained form of ordination). The amount of variance being accounted for in the CCA is the constrained inertia = 0.6976. So when you consider 0.6976 / 10.8656 = 6.42% of the variance being explained, that’s very poor. The quotient resulting from 0.6976 / 10.8656 is a familiar statistic, the coefficient of determination, r2.

You can use inertcomp() to compare the inertia of the constrained (CCA column) and unconstrained (CA column) variables to evaluate each to determine whether any contribute more than others:

inertcomp(cca.1, display="sites")

CCA CA

bcnp\_\_1 6.124480e-03 0.022534501

bcnp\_\_2 3.109255e-03 0.105137473

bcnp\_\_3 4.401853e-03 0.038878651

bcnp\_\_4 3.764973e-03 0.016896540

Etc.

inertcomp(cca.1, display="species")

CCA CA

junost 1.921746e-03 0.04300307

ameuta 2.527939e-03 0.06885333

arcpat 3.195900e-02 0.15614274

arttri 6.115870e-02 0.14465969

atrcan 4.091309e-02 0.12495667

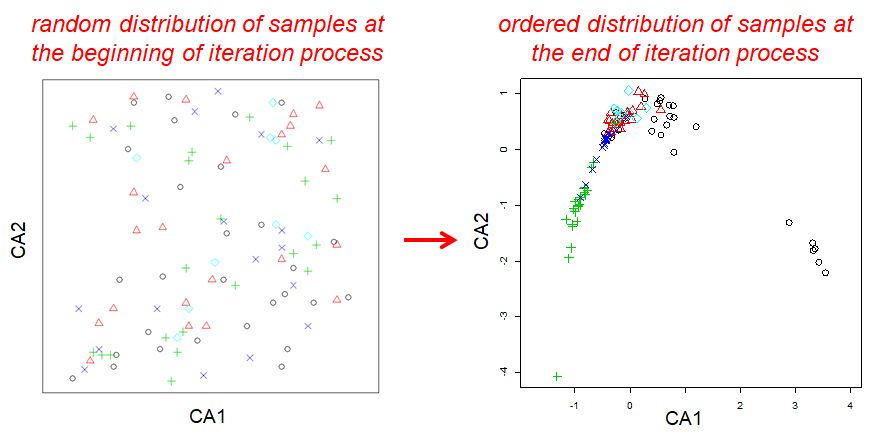
In all of the sites and species listed above, the unconstrained (CA) variance is very much larger than the constrained (CCA) variance, meaning once again that this CCA is poor. Ideally, you’d like the CCA >> CA.

The ratio of the CCA inertia to the total inertia is a measure of goodness of fit of your ordination:

goodness(cca.1, choices=1:4, statistic="explained", display="sites")

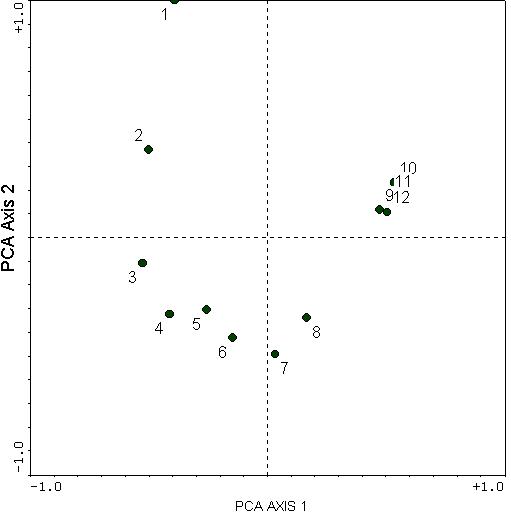
(You can also display species’ ratios with display="species".)

Sometimes, an **arch effect** can generate poor fit. An arch effect is a distortion in an ordination plot in which the second axis is an arched function of the first axis. It is caused by the unimodal distribution of species along gradients. (You may remember a horseshoe effect common in PCA; it is similar to an arch and is generated by similar non-independence, but a horseshoe has curved ends.)



An example of an arch effect in a correspondence analysis, from <https://www.davidzeleny.net/anadat-r/doku.php/en:ca_dca>

In a horseshoe, sites with no species in common appear very close to each other in the ordination diagram, which can happen for ordinations applied on datasets with many 0s:



An example of a horsehoe effect in a principal components analysis, from <http://ordination.okstate.edu/PCA.htm>

The toe can point up or down.

Correspondence analysis is susceptible to arch effect whereas CCA is not, but sometimes it happens even in CCA. If you re-run your CCA and include your variables squared in addition to untransformed variables and the ratio between CCA and CA improves, that is often indicative of an arch effect. But all too often, poor fit is indicative of explanatory factors that were not included in our analysis.

A permutation test (analysis of variance) can be used to examine model significance:

A <- anova(cca.1, test = "F")

A

Permutation test for cca under reduced model

Permutation: free

Number of permutations: 999

Model: cca(formula = veg ~ elev + slope + av)

Df ChiSquare F Pr(>F)

Model 3 0.6975 3.5673 0.001 \*\*\*

Residual 156 10.1680

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

This model is statistically significant, largely due to sample size (note the large number of residual degrees of freedom). This illustrates how something can be statistically significant yet with poor ecological significance (only 6.42% of the total variability was captured in the CCA [very low r2]). Even though this CCA is not doing a good job, once again I’ll continue with it to show you how to assess the effects of the environmental variables we chose to examine:

B <- anova(cca.1, by="terms")

B

Permutation test for cca under reduced model

Terms added sequentially (first to last)

Permutation: free

Number of permutations: 999

Model: cca(formula = veg ~ elev + slope + av)

Df ChiSquare F Pr(>F)

elev 1 0.5050 7.7484 0.001 \*\*\*

slope 1 0.1153 1.7691 0.026 \*

av 1 0.0772 1.1845 0.152

Residual 156 10.1680

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

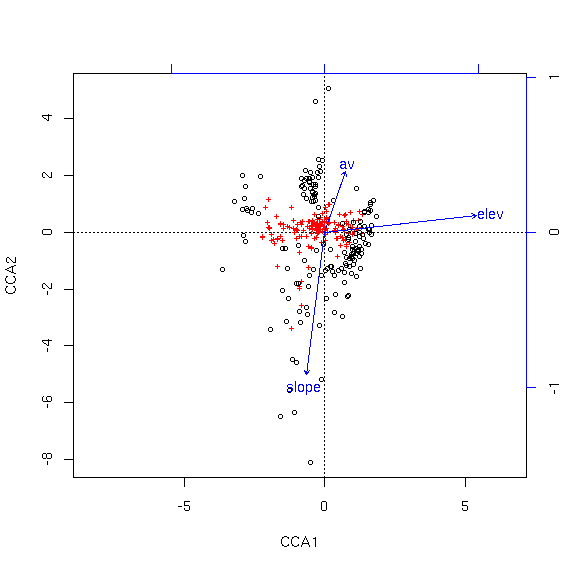
**Plotting the CCA:**

CCA generates plots that display species (in the plots to follow, these will be red crosses), sites (black circles), and explanatory environmental variables (blue arrows; arrows are used because these variables were numerical and thus are vectors). These plots are thus typically called **triplots** because they display three sets of data. Because they are displaying a lot of data on a single plot, their interpretation can be challenging, especially if variables are correlated with each other.

To plot your results, use the plot() function as follows:

cca1.plot <- plot(cca.1,choices=c(1,2))

*vegan* uses a convention where the plot() function not only produces the graphical plot but also stores information about the scaling and axes in a new object called an "ordiplot" for future use IF we create a new R object for it (in our case, cca1.plot). The choices=c(1,2) are the axes to plot. Since 1 and 2 are the defaults, they could be omitted.



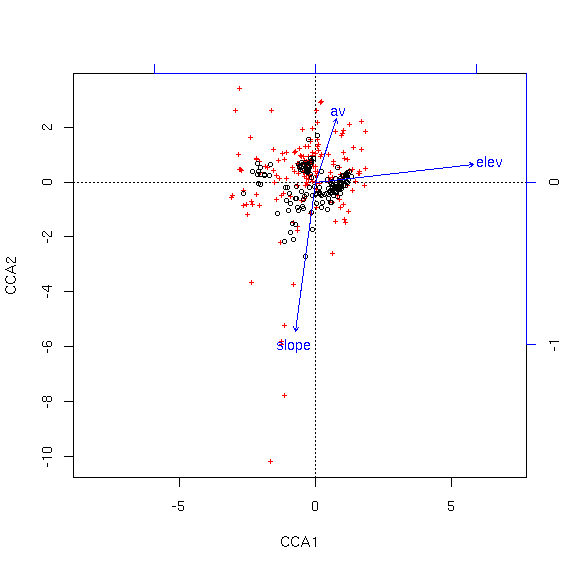
(You may have to stretch your plot window to see the ends of all the arrows.)

In a triplot, distances between object indicate their similarity. The angles between arrows represent associations (correlations) between environmental variables. 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. (Recall that av was not statistically significant, and it is the shortest of the three arrows.) 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.

Arrows and axes jointly reflect the composition of the major environmental gradients in the data. In this example, the first axis is associated with increasing elevation, whereas the second axis is associated with decreasing slope and increasing aspect value.

As you can see, the species are mostly condensed in the center of the ordination plot. To get a better look, we can specify "scaling=1", which effectively zooms in to the center of the graph.

cca2.plot <- plot(cca.1,scaling=1)



You can also stretch your plotting window in RStudio to help with visualization.

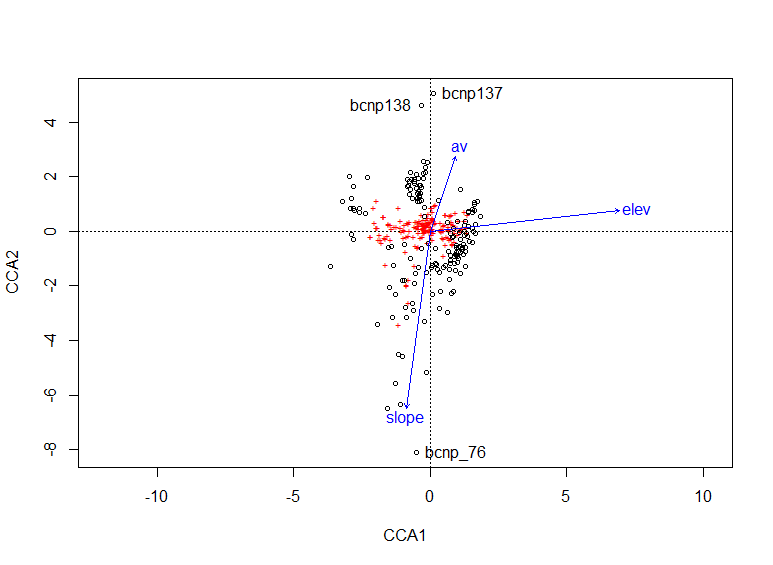
We can use the identify() function from *vegan* to identify specific sites in the cca1.plot triplot:

cca1.plot <- plot(cca.1)

identify(cca1.plot,what='sites')

Notice how the identify command uses 'sites' rather than 'plots' or 'samples' or some other synonym, and that the "ordiplot" object is what goes in as the first argument, not the calculated ordination.

For an example, I wanted to know what three of the “outermost” sites were:



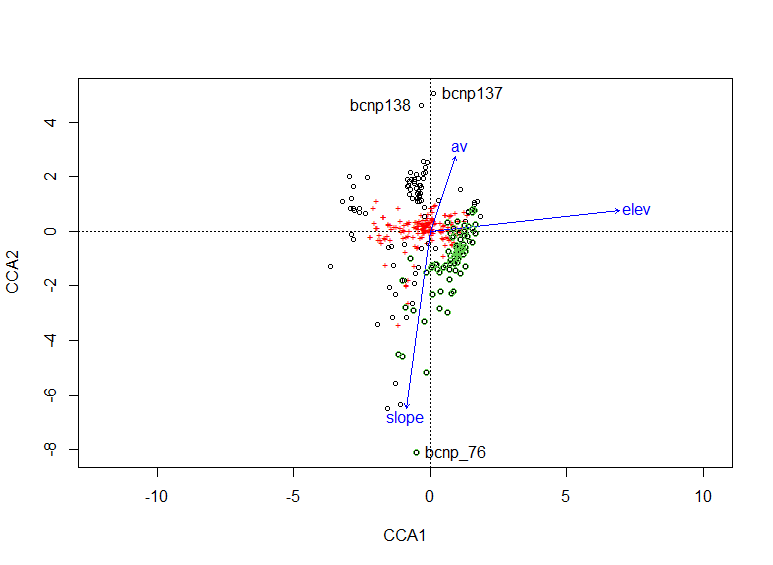
(Remember to hit the Esc key to stop identifying.)

You can see that they are sites 76, 137, and 138. Those sites could then be examined in more detail in future if desired.

You can also identify which sites contain a given species (such as arcpat = *Arctostaphylos patula*, green-leaf manzanita):

points(cca1.plot,what='sites',veg$arcpat>0,col=3)

The points() command identified points that meet a criterion, in this case sites that have arcpat present.

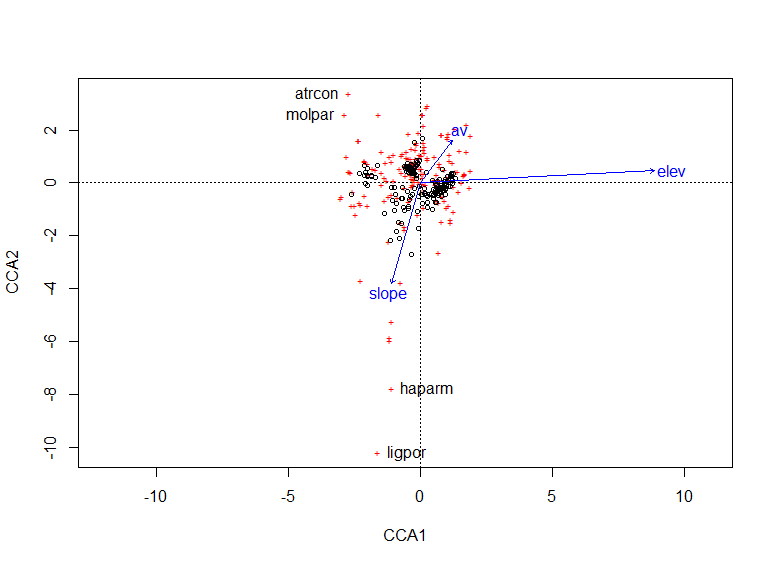


The circles in green in the resulting plot are ones with arcpat present.

Now let’s identify specific species on the plot. Let’s use the cca2.plot (the one that was “zoomed in”) for clarity:

cca2.plot <- plot(cca.1,scaling=1)

identify(cca2.plot,what='species')



I picked four of the outermost species to identify; these can then be examined in more detail if desired. (Additionally, a botanist with knowledge of the life history of these species may be able to make some inferences.)

**Adding categorical variables to the analysis:**

cca() allows you can add categorical variables to the formula you use in the weighted linear regression. As an example, we can add topographic position (which, if you examine site, has the following categories: bottom, low\_slope, mid\_slope, up\_slope, and ridge) to our model.

cca.3 <- cca(veg ~ elev + slope + av + pos)

cca.3

Call: cca(formula = veg ~ elev + slope + av + pos)

Inertia Proportion Rank

Total 10.8656 1.0000

Constrained 1.1994 0.1104 7

Unconstrained 9.6662 0.8896 147

Inertia is mean squared contingency coefficient

Eigenvalues for constrained axes:

CCA1 CCA2 CCA3 CCA4 CCA5 CCA6 CCA7

0.5268 0.2705 0.1292 0.0858 0.0723 0.0638 0.0510

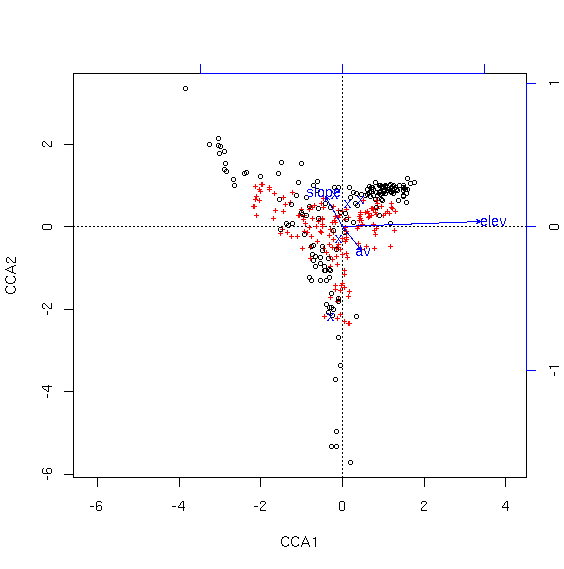
Eigenvalues for unconstrained axes:

CA1 CA2 CA3 CA4 CA5 CA6 CA7 CA8

0.6105 0.5365 0.4349 0.4074 0.3525 0.3274 0.2727 0.2598

(Showed only 8 of all 147 unconstrained eigenvalues)

plot(cca.3)



Notice how different this plot is from the first one (cca1.plot). Although the total variability explained did not increase very much (and it can't go down with an increase of degrees of freedom), regressing the vegetation against topographic position in addition to the other variables resulted in a quite different perspective on the distribution of the variability.

Each possible topographic position is plotted at the centroid of the samples of that type, shown as 5 blue Xs (because there were 5 topographic position categories). To find out which one is which, look at last element of the summary of the cca object:

summary(cca.3)

[previous lines of output, keep scrolling down the output]

Centroids for factor constraints

CCA1 CCA2 CCA3 CCA4 CCA5 CCA6

posbottom -0.27212 -2.2070 -0.2082 -0.6002 0.7309 0.08555

poslow\_slope -0.08944 -0.3111 0.7894 0.7438 -1.4639 -0.11783

posmid\_slope 0.13114 0.5434 -0.1861 0.7238 0.7434 0.33647

posridge 0.45090 0.6428 0.6320 -0.6593 0.8699 -0.66962

posup\_slope -0.19154 0.7557 -0.8329 -1.6023 -0.7748 -0.23554

Plots on topographic bottoms are relatively low on axis 1 and especially axis 2; ridge tops are higher on 1 and 2. Given their signs, bottoms and low\_slopes are qualitatively similar, mid\_slopes and ridges are similar, and up\_slopes are unique.

**Summary of ordination:**

We have now done four types of ordination: PCA, RDA, NMDS, and CCA. We did not cover correspondence analysis, detrended correspondence analysis, or a few other (not as commonly used) forms of ordination. The choice of which ordination to use should be made based on what your objectives are and what kind of data you have. Specifically:

PCA and NMDS are unconstrained whereas RDA and CCA are constrained.

PCA and RDA assume linear species responses whereas NMDS and CCA assume unimodal responses.

You use constrained ordinations to examine site-species-environment relationships (i.e., determine which environmental variables are associated with species occurrences/abundances at different sites), but even unconstrained ordinations can be used for that purpose in a two-step fashion (do an unconstrained ordination and then use the resulting ordinated axes as dependent variables in a second ordination that includes environmental variables). (In constrained ordinations, this is all done simultaneously.)

PCA is very commonly used and even more commonly mis-used. RDA is less common than CCA (likely because of its assumption of linear responses of species to environmental variables). NMDS and CCA are both very common. If you were to do an NMDS and a CCA on the same data, the NMDS should result in lower correlations between the key environmental variables and the ordination scores but should provide a better representation of overall community structure.

**References:**

McGarigal, K., S. Cushman, and S. Stafford. 2000. *Multivariate Statistics for Wildlife and Ecology Research*. Springer, New York, NY.

Palmer, M.W. 1993. Putting things in even better order: The advantages of canonical correspondence analysis. Ecology 74:2215-2230.

**Assignment:** due 0800 Monday, April 19

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.

Read in the grassland.community.csv site x species file from the course website as an object named comm (with header = TRUE, row.names = 1) and its associated site x environment file (plot.metadata.csv) as an object named metadata (with header = TRUE, row.names = 1). Do not use attach(metadata) for this assignment so as to avoid problems.

Examine the plot.metadata.csv file: there are three types of data represented. Habitat and site are text data (and are redundant since habitat type and site name are represented in each plot name); slope and aspect are numerical data; and slope.position and rel.moisture are categories. Slope and aspect are measured in degrees. Slope position and relative moisture are categories that could have been represented just as text (e.g. rel.moisture has values of 1-3 that represent dry, medium, and wet). Because they are represented as numbers, R will treat them as such even though doing so they are not truly numeric data.

**Q1. For the two true numeric variables (slope, aspect), construct a CCA. What % of the total variability was explained by this CCA? Use a permutation test (analysis of variance) to examine the significance of the model and of the variables; interpret the output. Then plot the ordination (first two axes) and interpret it.**

(For an example of how dangerous R can be, you can go ahead and do a CCA of the two categorical variables, and it will explain more variability than the CCA you did in Q1 and it will be statistically significant!)

Make an RMarkdown Word file of your work and turn that in. Be sure to include your answers to the 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 CCA