Lab 11: RDA and Variance partitioning

The goal of this lab is to apply Redundancy Analysis (RDA) to determine the influence of explanatory variables on patterns of variation in multivariate response variables. Constrained ordination is an extension of unconstrained ordination techniques in which the solution is constrained to be expressed by explanatory variables. You will also explore variable selection and variance partitioning to further interpret the results of RDA. These methods can also be applied to CCA.

# Set up R session

## Download packages

We will be using the following packages:

library(vegan)

## Import Data

Today you will be using the [Doubs fish data](https://search.r-project.org/CRAN/refmans/codep/html/Doubs.html) set (Verneaux 1973) from the Doubs River in France and Switzerland. This data set includes a *site by species* matrix (“DoubsSpe.csv”) that includes fish abundances for 27 species across 30 sites and a site by environment (“DoubsEnv.csv”) matrix containing measurements of 11 environmental variables across 30 sites.

spe <- read.csv("G:/Shared drives/MultivariateStatistics/Data/LabData/Lab11/Doubsspe.csv",  
 header = TRUE)  
  
env <- read.csv("G:/Shared drives/MultivariateStatistics/Data/LabData/Lab11/Doubsenv.csv",  
 header = TRUE)

Thus far, we have not encountered a categorical variable in an ordination (aside from binomial presence/absence data). Today we will recode the slope variable (pen) into a categorical variable (to explore how these are handled in the ordinations). Here you divide the data in the pen column into quantiles and assign a factor level to each quantile (very steep, steep, moderate, low) and place those into a new vector (pen2 that we will replace the original continuous slope variable, pen, with):

pen2 <- rep("very\_steep", nrow(env))  
pen2[env$pen <= quantile(env$pen)[4]] <- "steep"  
pen2[env$pen <= quantile(env$pen)[3]] <- "moderate"  
pen2[env$pen <= quantile(env$pen)[2]] <- "low"  
pen2 <- factor(pen2, levels = c("low", "moderate", "steep", "very\_steep"))  
table(pen2)

## pen2  
## low moderate steep very\_steep   
## 8 8 6 7

Incorporate the new qualitative slope variable “pen2” into the new environmental data frame (env2):

env2 <- env  
env2$pen <- pen2

# Data transformation

A common transformation for species data in RDA is the Hellinger transformation. This transformation is simply the square root of row normalized data (i.e., relative abundance). This transformation has been shown to have good statistical properties for constrained ordination. You will use the function decostand in the *vegan* package to perform this transformation:

spe.hel <- decostand(spe, "hellinger")

# Check for linear vs. unimodal response to determine whether to run an RDA vs. CCA

First, use Detrended Correspondence Analysis (DCA) to determine the length of the canonical axes. You will use the decorana function in the *vegan* Library. While DCA is a separate analysis with its own assumptions and multifaceted output, you will focus on axis length. An axis length > 3 is evidence of a unimodal relationship. An axis length of <3 is evidence of a linear relationship.

decorana(spe.hel)

Let’s check out the response relative to the first axis with a CCA like we did in Lab 10:

sp.CCA <- cca(spe.hel ~ ., data = env2)

Function for plotting species abundances vs. CCA Axis 1:

f2 <- function(x) {  
 plot(x ~ sp.CCA$CC$wa[, 1], xlab = "CCA AXIS 1", ylab = "Abundance ")  
}  
  
# Apply the function across all the species:  
  
mapply(f2, spe.hel)

# Question 1: Based on the axis lengths and graphs, do you think a linear (RDA) or unimodal (CCA) response model is appropriate? (10 pts)

Unfortunately, none of these criteria are entirely satisfying as they seem plagued by the problem of circularity. Specifically, the DCA and CCA models assume unimodal species responses and fits the data accordingly. For these reasons, we are going to pursue this analysis with a linear model (RDA), so that we can explore a different type of model this week. If you were doing this analysis for your own research, you would want to do your best to justify the choice between CCA and RDA. Once again this can be an instance where the decisions made during an analysis seem like more of an art than a science.

# Run RDA using function rda in vegan using all the environmental variables.

`?`(rda)

spe.rda <- rda(spe.hel ~ ., env2)

Explore the results:

summary(spe.rda)

Unadjusted R^2 retrieved from the rda object

R2 <- RsquareAdj(spe.rda)$r.squared  
R2

## [1] 0.7270878

Adjusted R^2 retrieved from the rda object

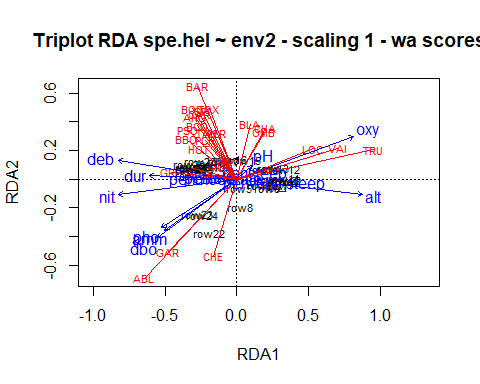
R2adj <- RsquareAdj(spe.rda)$adj.r.squared  
R2adj

## [1] 0.5224036

# Question 2: How many explanatory variables did we use in this fitted model? How might the adjusted R-squared change if we reduced the number of explanatory variables in the RDA? (10 pts)

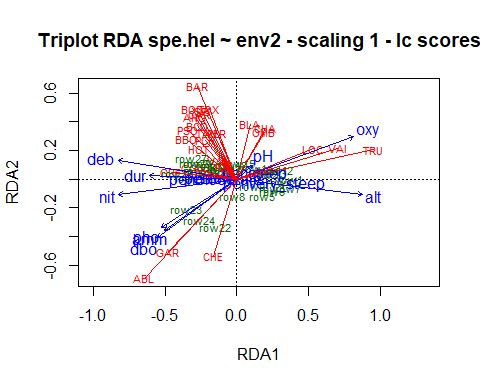
Plot using the F-scores:

plot(spe.rda, scaling = 1, main = "Triplot RDA spe.hel ~ env2 - scaling 1 - wa scores")  
spe.sc <- scores(spe.rda, choices = 1:2, scaling = 1, display = "sp")  
arrows(0, 0, spe.sc[, 1], spe.sc[, 2], length = 0, lty = 1, col = "red")



Plot using the Z-scores:

plot(spe.rda, scaling = 1, display = c("sp", "lc", "cn"), main = "Triplot RDA spe.hel ~ env2 - scaling 1 - lc scores")  
arrows(0, 0, spe.sc[, 1], spe.sc[, 2], length = 0, lty = 1, col = "red")



# Question 3: Why do the arrows for dbo, pho, and amm plot so closely together? (10 pts)

Conduct a permutation test using anova function in vegan to test the significance of the model and individual axes:

anova(spe.rda, step = 1000)

## Permutation test for rda under reduced model  
## Permutation: free  
## Number of permutations: 999  
##   
## Model: rda(formula = spe.hel ~ alt + pen + deb + pH + dur + pho + nit + amm + oxy + dbo, data = env2)  
## Df Variance F Pr(>F)   
## Model 12 0.36537 3.5522 0.001 \*\*\*  
## Residual 16 0.13714   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Tests of all canonical axes:

anova(spe.rda, by = "axis", step = 1000)

## Permutation test for rda under reduced model  
## Forward tests for axes  
## Permutation: free  
## Number of permutations: 999  
##   
## Model: rda(formula = spe.hel ~ alt + pen + deb + pH + dur + pho + nit + amm + oxy + dbo, data = env2)  
## Df Variance F Pr(>F)   
## RDA1 1 0.228081 26.6098 0.001 \*\*\*  
## RDA2 1 0.053696 6.2646 0.003 \*\*   
## RDA3 1 0.032124 3.7478 0.359   
## RDA4 1 0.023207 2.7075 0.781   
## RDA5 1 0.008707 1.0159 1.000   
## RDA6 1 0.007218 0.8421 1.000   
## RDA7 1 0.004862 0.5673 1.000   
## RDA8 1 0.002919 0.3406 1.000   
## RDA9 1 0.002141 0.2497 1.000   
## RDA10 1 0.001160 0.1353 1.000   
## RDA11 1 0.000913 0.1066 1.000   
## RDA12 1 0.000341 0.0397 1.000   
## Residual 16 0.137141   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Variable Selection: forward selection using ordiR2step in *vegan*.

Hopefully, my questions have led you to realize that we have used too many variables and that many variables are closely related (i.e., collinear) in our data set. Here we want to reduce the number of variables while maintaining a model that describes as much variance as possible. I have led you into this trap to provide an example where we use stepwise variable selection to select the predictor variables for our RDA. See settings in ordiR2step for more options.

`?`(`?`(ordiR2step))  
step.forward <- ordiR2step(rda(spe.hel ~ 1, data = env2), scope = formula(spe.rda),  
 R2scope = F, direction = "forward", pstep = 1000)

# Question 4: Describe the algorithm used by the forward selection process. How are variables selected for inclusion in the model? (20 pts)

# Question 5: What is the most parsimonious model? (10 pts)

# Question 6: Conduct an RDA with the most parsimonious model. (10 pts)

# Partial RDA

Partial RDA allows us to quantify the unique contribution of an explanatory variable in describing variation in a response data set. It allows us to hold other explanatory variables constant while assessing the role of the focal variable. Here, I show a partial RDA for the altitude variable “alt”. You can conduct a partial RDA for any or all explanatory variables.

partial.alt <- rda(spe.hel ~ alt + Condition(oxy + dbo + dur), data = env2)

# Question 7: Explore the results of this model and conduct a permutation test to assess the significance of this model. You will need to write some code to do so. (10 pts)

Feel free to explore partial RDA for other variables in the data set.

# Variance partitioning

Variance partitioning allows us quantify the amount of variation attributed to each variable and the amount that is shared between variables. In some ways, it is an extension of partial RDA. We can partition variance between up to four variables (or matrices) with the varpart function in *vegan*.

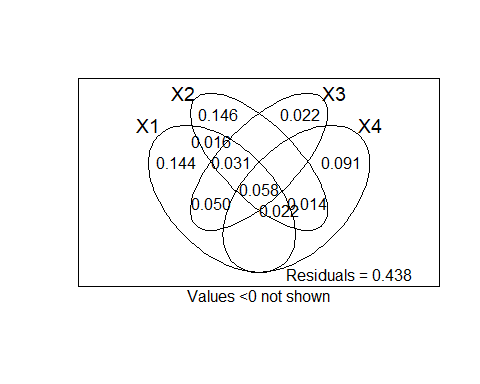
`?`(`?`(varpart))

Here we partition the variance for the model we constructed through forward selection above:

spe.part <- varpart(spe.hel, ~alt, ~oxy, ~dur, ~dbo, data = env2)  
spe.part

##   
## Partition of variance in RDA   
##   
## Call: varpart(Y = spe.hel, X = ~alt, ~oxy, ~dur, ~dbo, data = env2)  
##   
## Explanatory tables:  
## X1: ~alt  
## X2: ~oxy  
## X3: ~dur  
## X4: ~dbo   
##   
## No. of explanatory tables: 4   
## Total variation (SS): 14.07   
## Variance: 0.50251   
## No. of observations: 29   
##   
## Partition table:  
## Df R.square Adj.R.square Testable  
## [aeghklno] = X1 1 0.32807 0.30318 TRUE  
## [befiklmo] = X2 1 0.30248 0.27665 TRUE  
## [cfgjlmno] = X3 1 0.17349 0.14287 TRUE  
## [dhijkmno] = X4 1 0.19244 0.16253 TRUE  
## [abefghiklmno] = X1+X2 2 0.49209 0.45302 TRUE  
## [acefghjklmno] = X1+X3 2 0.36084 0.31167 TRUE  
## [adeghijklmno] = X1+X4 2 0.44604 0.40343 TRUE  
## [bcefgijklmno] = X2+X3 2 0.38730 0.34017 TRUE  
## [bdefhijklmno] = X2+X4 2 0.39263 0.34591 TRUE  
## [cdfghijklmno] = X3+X4 2 0.31007 0.25699 TRUE  
## [abcefghijklmno] = X1+X2+X3 3 0.52791 0.47125 TRUE  
## [abdefghijklmno] = X1+X2+X4 3 0.58942 0.54016 TRUE  
## [acdefghijklmno] = X1+X3+X4 3 0.47904 0.41653 TRUE  
## [bcdefghijklmno] = X2+X3+X4 3 0.48052 0.41818 TRUE  
## [abcdefghijklmno] = All 4 0.62469 0.56214 TRUE  
## Individual fractions   
## [a] = X1 | X2+X3+X4 1 0.14396 TRUE  
## [b] = X2 | X1+X3+X4 1 0.14561 TRUE  
## [c] = X3 | X1+X2+X4 1 0.02199 TRUE  
## [d] = X4 | X1+X2+X3 1 0.09089 TRUE  
## [e] 0 0.01557 FALSE  
## [f] 0 -0.00889 FALSE  
## [g] 0 0.05028 FALSE  
## [h] 0 -0.01287 FALSE  
## [i] 0 0.01397 FALSE  
## [j] 0 -0.00376 FALSE  
## [k] 0 0.02214 FALSE  
## [l] 0 0.03108 FALSE  
## [m] 0 -0.00084 FALSE  
## [n] 0 -0.00499 FALSE  
## [o] 0 0.05800 FALSE  
## [p] = Residuals 0 0.43786 FALSE  
## Controlling 2 tables X   
## [ae] = X1 | X3+X4 1 0.15954 TRUE  
## [ag] = X1 | X2+X4 1 0.19425 TRUE  
## [ah] = X1 | X2+X3 1 0.13109 TRUE  
## [be] = X2 | X3+X4 1 0.16119 TRUE  
## [bf] = X2 | X1+X4 1 0.13672 TRUE  
## [bi] = X2 | X1+X3 1 0.15958 TRUE  
## [cf] = X3 | X1+X4 1 0.01310 TRUE  
## [cg] = X3 | X2+X4 1 0.07227 TRUE  
## [cj] = X3 | X1+X2 1 0.01823 TRUE  
## [dh] = X4 | X2+X3 1 0.07802 TRUE  
## [di] = X4 | X1+X3 1 0.10486 TRUE  
## [dj] = X4 | X1+X2 1 0.08713 TRUE  
## Controlling 1 table X   
## [aghn] = X1 | X2 1 0.17638 TRUE  
## [aehk] = X1 | X3 1 0.16880 TRUE  
## [aegl] = X1 | X4 1 0.24090 TRUE  
## [bfim] = X2 | X1 1 0.14985 TRUE  
## [beik] = X2 | X3 1 0.19729 TRUE  
## [befl] = X2 | X4 1 0.18338 TRUE  
## [cfjm] = X3 | X1 1 0.00850 TRUE  
## [cgjn] = X3 | X2 1 0.06352 TRUE  
## [cfgl] = X3 | X4 1 0.09446 TRUE  
## [dijm] = X4 | X1 1 0.10025 TRUE  
## [dhjn] = X4 | X2 1 0.06926 TRUE  
## [dhik] = X4 | X3 1 0.11412 TRUE  
## ---  
## Use function 'rda' to test significance of fractions of interest

plot(spe.part, digits = 2)



# Question 8: Would these analyses be appropriate for your own data set? (20 pts)