Variance Partitioning and Commonality Analysis

Mariah Casmey

Video seminar: http://tinyurl.com/mv690/seminars/vpart
Data: http://tinyurl.com/mv690/seminars/vpart/data

Introduction

One issue that is encountered in methods with multiple predictor variables is understanding the contributions of these variables to explaining the variation in the response variable. From plots we can see the effects of predictor variables on the response variable but distinguishing the effects of one predictor from another can be difficult. This is especially true in the cases of multicollinearity where predictor variables are linearly related to one another. Enter variance partitioning which gives us a way to break down the variance explained by predictor variables in redundancy analysis (RDA), canonical correspondence analysis (CCA), and multiple linear regression analysis. This can be used to quantify the unique variance explained by individual predictor variables or predictor tables as well as their overlap with other variables in the model. In this way, variance partitioning enables us to better understand the effects of our predictor variables on the response variable.

Variance partitioning works by computing variance explained by variables in partial models. Partial models are models creating using a subset of the predictor variables in the full RDA / CCA / multiple linear model. The fractions of variance explained in these partial models can then be used to calculate the fraction of variance explained by each variable. In this method, we use the adjusted R^2 of each model as the metric of variance explained. We use the adjusted R^2 in place of unadjusted R^2 as values for different models will be inflated if there are a difference number of variables in a given model. This is because the value of R^2 will naturally increase as the number of variables in a model increases. At a certain point, additional variables stop explaining variance in the model and start to explain random effects, which causes the model to look like more variance is explained (a phenomena known as ‘overfitting’).

Note: Variance partitioning and commonality analysis refer to the same methodology and you will see people using them interchangeably. However, some people use variance partitioning to refer exclusively to when it is used for RDA and CCA whereas they would use commonality analysis when it is used for univariate multiple linear regression.

Variance Partitioning for RDA and CCA

In this first example we will manually partition the variance for a dataset to illustrate how this method works. Then we will verify these results using the `varpart` function from the vegan library

```
temp <- read.csv("./AB_Climate_Trees.csv")
head(temp)
```

```r
rownames(temp) <- temp$ECOSYS #Set row names
ecolabels=temp$ECOSYS
species.dat=temp[,11:23] #extract species frequent to its own data frame
environ.dat=temp[,3:10] #extract climate data to its own data frame
geographic.dat=temp[,2] #extract location data

library(vegan)

#Let's look at how all of our data looks
edaPlot <- rda(species.dat ~ ., data=environ.dat)
plot(edaPlot, choices=c(1,2), type="text")

# Notice how MCMT and MAT seem to be somewhat related. Let's investigate with
variance partitioning
# We will create a model using only these two variables to examine their
relationship with the species frequency data.

# RDA of full model, gives us the fractions of [a+b+c]
rda.all <- rda(species.dat ~ MAT + MCMT, data=environ.dat)
# fractions [a+b]
rda.MAT <- rda(species.dat ~ MAT, data = environ.dat)
#fractions [b+c]
rda.MCMT <- rda(species.dat ~ MCMT, data = environ.dat)

plot(rda.all, choices=c(1,2), type="text") #check the plot

#fractions [a+b+c]
RsquareAdj(rda.all)

abc <- RsquareAdj(rda.all)$adj.r.squared # Extract the adjusted r-squared for
the full model

#fraction [a+b]
ab <- RsquareAdj(rda.MAT)$adj.r.squared

#fraction [b+c]
bc <-  RsquareAdj(rda.MCMT)$adj.r.squared

#individual fractions
b <- ab + bc - abc
a <- ab - b
c <- bc - b

out <- varpart(species.dat, ~ MAT, ~ MCMT, data=environ.dat)

#Let's see if our calculations match those that `varpart` calculated
out$part$fract #Fractions (ab, bc, abc)
```

```
out$part$indfract #Individual Fractions (a,b,c)

#The nice feature about varpart is that it allows you to plot your results as
a Venn diagram
plot(out, bg = c("hotpink","skyblue"), Xnames = c("MAT", "MCMT"))
```

In this example we only manually calculated the fractions for a model with two predictor variables. Doing this kind of analysis can be helpful in better understanding how your explanatory variables relate to one another. Consider what the plot you just created tells you about the relationship of MCMT and MAT. Can you tie this back to what you saw in the first RDA plot you made (hint: think about what vector length signifies).

Though we only investigated two variables in the example, `varpart` can handle up to four explanatory variables or tables. This can be helpful if you have several variables that all appear to be related to one another (from exploratory data analysis) or as indicated by the literature.

The overall variance explained by the variable will always be positive. but sometimes a fractions of a variable will be negative. At first this seems impossible - how can a variable explain negative variance? But what this is telling us that the variable is actually interfering the variance explained in the rest of the model (Peterson and Mahajan 1976).

Significance testing of variance partitioning

We can also calculate p-values for the variation fractions we calculate by using permutations of the R^2 values for each fraction. The output of `varpart` contains a *Testable* column which indicates whether we can express the fraction as an RDA for testing. We can construct an RDA for each fraction and calculate p-values using the `anova` function. It is important that the full model be significant before individual partitions should be examined. In this case, a significant p-value indicates. This process is demonstrated below using the previous model.

```
#Which fractions can we test?
out <- varpart(species.dat, ~ MAT, ~ MCMT, data=environ.dat)
out #The only one we can't test is [b]

# RDA of full model, gives us the fractions of [a+b+c]
rda.all <- rda(species.dat ~ MAT + MCMT, data=environ.dat)
# Marginal effect of MAT: fractions [a+b]
rda.MAT <- rda(species.dat ~ MAT, data = environ.dat)
# Marginal effect of MCMT: fractions [b+c]
rda.MCMT <- rda(species.dat ~ MCMT, data = environ.dat)
# Partial effect of MAT [a]
rda.MAT.MCMT <- rda(species.dat~ MAT + Condition(MCMT), data = environ.dat)
# Partial effect of MCMT [c]
rda.MCMT.MAT <- rda(species.dat ~ MCMT + Condition(MAT), data = environ.dat)

anova(rda.all)      #[a+b+c]
anova(rda.MAT)      #[a+b]
```

```r
anova(rda.MCMT)     #[b+c]
anova(rda.MAT.MCMT) #[a]
anova(rda.MCMT.MAT) #[c]
```

Partitioning variance between different tables of variables

With this varpart function that we have just confirm works as expected, we can also all investigate the variance explained by different types of variables. This was the original purpose of variance partitioning in community ecology as described by Borcard et al. 1992. They were interested in partitioning out the variance explained by geographic gradients from environmental variables. Below we work through some simulated data to see how much variance is uniquely explained by environment variables for imaginary species.

```r
temp <- read.csv("./SpeciesDat.csv")
head(temp) # Check data

rownames(temp) <- paste(temp$BIOME, temp$plotNum)
species.dat=temp[,10:15] #Extract species columns
environ.dat=temp[,c(4:6, 9)] #Extract environmental variables
geo.dat=temp[,2:3] #Extract geographical variables (longitude and latitude in
our case)

#Before we do any analysis, look at ordination plots.

#Plot all the variables of interest, do any variables seem highly related to
the geographic variables?
eda.both <- rda(species.dat, temp[,c(2:6, 9)])
plot(eda.both, choices=c(1,2), type="text")

#Let's see how our just environmental data looks
eda.environ <- rda(species.dat ~ ., data=environ.dat)
plot(eda.environ, choices=c(1,2), type="text")
eda.environ

#Let's also see how things look just based on latitude and longitude.
#Does it look like some variance is explained just by geographical gradients?
eda.geo <- rda(species.dat ~ ., data=geo.dat)
plot(eda.geo, choices=c(1,2), type="text")

#Variance partitioning
out <- varpart(species.dat, environ.dat, geo.dat)
plot(out, bg = c("hotpink","skyblue"), Xnames = c("Environmental Data",
"Geographical Data"))
#Does it look like the geographical data explains a significant amount of
variance in the data beyond the environmental data?

# Let's test that with significance testing
out #This tells us ab, bc, abc, a, and c are testable. b is not
```

```
anova(eda.both) #abc
anova(eda.environ) #ab
anova(eda.geo) #bc
anova(rda(species.dat ~ environ.dat$MAT + environ.dat$MSP + environ.dat$ELEV
+ environ.dat$MCMT + Condition(geo.dat$lat_y + geo.dat$long_x))) #a
anova(rda(species.dat ~ environ.dat$MAT + environ.dat$MSP + environ.dat$ELEV
+ environ.dat$MCMT + Condition(geo.dat$lat_y + geo.dat$long_x))) #c
```

Variance Partitioning for Multiple Linear Regression

Variance partitioning can also be done for univariate multiple linear regression analysis. I will not go too much into multiple linear regression analysis as it is not the topic of this seminar nor covered in this course, but if you know how to use it then this could be a helpful tool. In the following example we use a dataset from the `lattice` library to first build a multiple linear model for predicting ozone levels and then we partition the variance of that model.

```
ozone <- lattice::environmental
head(ozone)
plot(ozone)

#Looking at the plots, temperature and wind seem to have a negative linear
correlation

lm.oz <- lm(ozone~ radiation + temperature + wind, data=ozone)
summary(lm.oz)
anova(lm.oz) #Temperature seems to be a good predictor variable based on sum
sq.
plot(lm.oz) #The curvature in the residuals plot is a little concerning, but
let's proceed

#For the linear model we add the variables as so: varpart(Y, X1, X2, X3)
varp.oz <- varpart(ozone$ozone, ozone$radiation, ozone$temperature,
ozone$wind)
plot(varp.oz, bg = c("red", "yellow", "blue"), Xnames = c("Radiation",
"Temperature", "Wind"))
```

As you can see the syntax is quite similar for the univariate and multivariate cases, it just depends on the data you provide the `varpart` method.

Optional Including Interaction terms

An interaction term actually improves the model as shown below. We can also examine how the variance explained by this interaction term .

```
library(car)
library(ggplot2)
ozone$wind_cut <- cut_number(ozone$wind, 4)
ggplot(ozone, aes(x = temperature, y = ozone)) + geom_point() +
facet_wrap(~wind_cut, nrow=2) #Looks like there is some interaction between
```

```
wind and temperature

lm.oz2 <- lm(ozone ~ temperature * wind + radiation, data = ozone)
summary(lm.oz2)
anova(lm.oz2) #Interaction term is significant
plot(lm.oz2)

varp.oz2 <- varpart(ozone$ozone, ozone$temperature * ozone$wind,
ozone$temperature, ozone$wind, ozone$radiation)
varp.oz2
plot(varp.oz2, bg = c("#003f5c", "#bc5090", "#ff6361", "#ffa600"), Xnames =
c("Temp * Wind", "Temp", "Wind", "Rad"))
```

References and Further Reading

Borcard, D., P. Legendre, and P. Drapeau. 1992. Partialling out the Spatial Component of Ecological Variation. Ecology 73:1045–1055.

Økland, R. H., and O. Eilertsen. 1994. Canonical Correspondence Analysis with Variation Partitioning: Some Comments and an Application. Journal of Vegetation Science 5:117–126.

Peterson, R. A., and V. Mahajan. 1976. Practical Significance and Partitioning Variance in Discriminant Analysis. Decision Sciences 7:649–658.

Zeleny, D. (n.d.). Variation partitioning (constrained ordination). https://www.davidzeleny.net/anadat-r/doku.php/en:varpart.