Multivariate statistics

Vegan, Diversity & Dissimilarities

Francesco Maria Sabatini

2021-01-19

Code adapted from the vegan tutorial, Oksanen 2015. Additional acknowledgements to B. Jimenez-Alfaro & O. Purschke

```
# download packages, if necessary
install.packages(c("vegan", "FD", "psych", "pheatmap"))
```

```
#load packages
library(vegan)
library(FD)
library(psych) #for pairs-plot
library(pheatmap) #for heatmaps
```

PART 1 - INTRODUCTION TO VEGAN

Vegan comes with three datasets

```
# Take a look at vegan's vignettes
browseVignettes("vegan")
```

*Dataset 1 - **varespec***
From Vaere et al (1995) Journal of Vegetation Science 6, 523-530.
Data on floristic composition (estimated cover in %) and environmental variables in floristically homogeneous oligotrophic Pinus sylvestris forests with heath-like understorey vegetation in Fennoscandia to study the effect of lichen grazing by reindeer on the understorey vegetation

```
data(varespec)
data(varechem)
?varespec
```

*Dataset 2 - **BCI***
From Condit et al (2002) Science 295: 666-669 Tree counts in 1-hectare plots in the Barro Colorado Island.

```
data(BCI)
data(BCI.env)
?BCI
```

*Dataset 3 - **dune***
From Jongman et al (1987) Data Analysis in Community and Landscape Ecology. Wageningen
vegetation data (2x2m plots) from dune meadows (cover-abundance scale of 9-degree Braun-Blanquet) environmental data to estimate the effect of management

```
data(dune)
data(dune.env)
?dune
```

Q1 - Explore these datasets from vegan. For each pair of datasets answer the following questions

1. What kind of data represents each data.frame (species, environmental or trait data)?
2. How many plots, species, and environmental variables are there for each dataset?
3. What kind of variables are contained (binary, nominal, ordinal, quantitative) in each data.frame?

```
# tips
?str
?class
?summary
?head
?tail
?dim
?nrows
?ncols
?rownames
?colnames
?range
?apply  #e.g. apply(varespec, MARGIN=1, "max")
```

Q2 - Choose one dataset and find the function in vegan to

1. Convert species abundances to presence-absence (from quantitative to binary)
2. Scale species abundances between [0,1]
3. Scale species abundances by dividing them by site totals
4. Standardize environmental variables so that they are all centered on their mean, and are expressed in s.d. units

```
#Hint # Get help on the decostand() function
?decostand
```

```
# Transformation and standardization of the species data
## Simple transformations

# Partial view of the raw data (abundance codes)
varespec[1:5, 2:4]

## 1) Transform abundances to presence-absence (1-0)
varespec.pa <- decostand(varespec, method = "pa")
varespec.pa[1:5, 2:4]

## 2) Standardization by columns (species)
# Scale abundances by dividing them by the maximum value of each
# species
# Note: MARGIN = 2 (column, default value) for argument "max"
varespec.scal <- decostand(varespec, "max")
varespec.scal[1:5, 2:4]
# Display the maximum in each transformed column
apply(varespec.scal, 2, max)

## 3) Standardization by rows (sites)
# Scale abundances by dividing them by the site totals
# (profiles of relative abundance by site)
varespec.rel <- decostand(varespec, "total") # default MARGIN = 1
varespec.rel[1:5, 2:4]
# Display the sum of row vectors to determine if the scaling worked
# properly
```

```r
rowSums(varespec.rel) # equivalent to: apply(varespec.rel, 1, sum)

## 4) Standardization to zero mean and unit s.d.
varechem.st <- decostand(varechem, "standardize")
# verify it worked
colMeans(varechem.st)
apply(varechem.st, MARGIN=2, sd)
```

PART 2 - DIVERSITY ANALYSIS

Q3 - Focus on the BCI dataset and

1. Derive the most common diversity metrics (species richness per plot, total shannon per plot, simpson per plot). Build an histogram for each
2. Build an individual-based rarefaction curve for each plot
3. Build a sample-based rarefaction curve for the whole dataset. What's the conceptual difference with the previous?

```r
#Hint # Get help on the following functions in vegan
?diversity
?specaccum
?rarefy
?rarecurve
```

```r
# Species richness
specnumber(BCI) # returns the species richness per plot
summary(specnumber(BCI)) # report descriptive statistics
hist(specnumber(BCI)) # have a look to the frequencies

# Diversity indices
diversity(BCI, index = "shannon") # Computes shannon diversity index per plot
hist(diversity(BCI, index = "shannon")) # see the histogram
diversity(BCI, index = "simpson") # Computes shannon diversity index per plot
hist(diversity(BCI, index = "simpson")) # see the histogram

# rarefaction (individual-based rarefaction)
rarefy(BCI, 20) # gives you the species per 20 individuals
rarecurve(BCI) # sample size reflects individuals

# rarefaction (sample-based)
spa <- specaccum(BCI)
plot(spa) # Plot the rarefaction curve
plot(spa, ci.type="poly", col="blue", lwd=2, ci.lty=0, ci.col="lightblue") # just nicer
```

PART 3 - ECOLOGICAL DISTANCES

Q4 - Find the function in vegan to calculate dissimilarity matrices

1. When doing an analysis of Q mode (comparing-objects), what dissimilarity measures could I use for species data? [remember the double-zero problem]. Which measure would you use for an R mode analysis instead (between-variables)?
2. What type of object (=class) is a distance matrix? What dimensions does it have? How can I convert it to a matrix?
3. How many dissimilarity matrices can I potentially calculate (with the same dissimilarity measure) when both species and environmental data are available?

4. Explore the differences between dissimilarity matrices (Q-mode), when using different metrics, and between species and environmental data. To what extent do they match? It can be done graphically with the function `pheatmap` (from the homonymous package) or quantitatively using the Mantel test
5. Find the most similar (less dissimilar) couple of plots in your dataset, based on species data. Is the same couple also the most similar when considering the environmental variables?

Information on dissimilarity measures are available at: https://www.davidzeleny.net/anadat-r/doku.php/en:div-ind

```
#Hints #
?vegdist
?as.matrix
?vegan::mantel #!! the FD package masks the mantel function from vegan!
```

```
## 1)
distance.spe.q1 <- vegdist(varespec)
distance.spe.q2 <- vegdist(varespec, method ="euclidean") #does not account for double 0s!
distance.spe.q3 <- vegdist(varespec, binary=FALSE)
# binary = FALSE looks at the abundance;
distance.spe.q4 <- vegdist(varespec, binary=TRUE)
# TRUE looks at presence-absence (Sorenson's index)
# equivalent to distancce.spe.q1

## 2)
dim(distance.spe.q1)
length(distance.spe.q1) #How do you obtain this number?
distances <- as.matrix(distance.spe.q1)
dim(distances)

## 3.1) Q mode analysis - dissimilarity between sites based on species data
distance.spe.q1 <- vegdist(varespec) # "bray" is the default, not necessary to type

## 3.2) Q mode analysis - dissimilarity between sites based on environmental variables
distance.env.q1 <- vegdist(varechem, "euclidean")
# even better would be to standardize the variables
varechem.st <- decostand(varechem, method = "standardize", MARGIN=2) #by column!
distance.env.q1 <- vegdist(varechem.st, "euclidean")

## 3.3) R mode analysis - dissimilarity between species,
## based on their presence absence in a site
## Need to Transpose - in R variables are conventionally in columns!
varespec.t <- t(varespec)
#transpose matrix of species abundancces
distance.spe.r1 <- dist(decostand(varespec, "chi.square"))
# dist = euclidean distance in {base}

## 3.4) R mode analysis -  between environmental predictors
distance.env.r1 <- vegdist(t(varechem.st), method ="euclidean")
# To compare env. variables it does not really make sense to calculate dissimilarities
# (even if technically possible). Tather one should calculate correlations
psych::pairs.panels(varechem[,1:5],
             method = "pearson", # correlation method
             hist.col = "#00AFBB",
             density = TRUE,  # show density plots
             ellipses = TRUE # show correlation ellipses
```

```
            )

## 4) Compare different dissimilarity matrices graphically
# dissimilarities based on species (Q-mode), with different metrics
pheatmap::pheatmap(distance.spe.q1, cluster_rows = F, cluster_cols = F)
pheatmap::pheatmap(distance.spe.q2, cluster_rows = F, cluster_cols = F)
pheatmap::pheatmap(distance.spe.q4, cluster_rows = F, cluster_cols = F)

# The Mantel test ####
# Mantel test calculates correlations between dissimilarities
# The higher is the result, the more similar the groups compared
vegan::mantel(distance.spe.q1, distance.spe.q2)

## 5) find the most similar plots
#transform to matrix, and replace diagonal with ones
#[to avoid considering distances of plots to themselves]
totest <- as.matrix(distance.spe.q1) + diag(nrow=nrow(varespec))
minn <- which(totest==min(totest), arr.ind=T)
totest[minn]
```

Additional question for Advanced users Write a short function to calculate jaccard dissimilarity between any two plots.
Djac = 1- a/(a+b+c)
Where a are the species in common between the two sites, b the species unique in site i, and c the species unique to site j

```
# my solution
myjac <- function(x,y){
  x.pa <- rbind(x,y)>0
  return(1 - sum(colSums(x.pa)==2) / (sum(colSums(x.pa)==2) + sum(colSums(x.pa)==1)))
  }
```