Multivariate statistics

Classification & Ordination

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"))
```

```
#load packages
library(vegan)
library(FD)
```

```
# clean workspace
rm(list=ls())
#reimport data
data(dune)
data(dune.env)
data(varespec)
data(varechem)
data(BCI)
data(BCI.env)
```

PART 4 - CLASSIFICATION

Q5 - Now, let's consider the dune dataset. You are asked to produce different classifications and compare them

We first have to choose a dissimilarity which is appropriate for our task.

1. Focus on species data, run a k-means classification (with 5 groups). Explore the output and then calculate the mean value of the 'A1' [thickness of soil A1 horizon, from the data.frame `dune.env`] for each group. [for advanced users - build a boxplot of A1 per group]

2. Classify again your data with the following hierarchical algorithms: single-linkage, complete-linkage, UPGMA (=average). Produce and compare the dendrograms for these three algorithms. How do they differ? Why? Compute again the means of A1 when dividing (=cutting) the dendrogram in 5 groups

3. Calculate a cophenetic matrix for each of your dendrograms (don't forget to explore the output!): how well are they correlated to your original dissimilarity matrix? Which algorithm better fits your data?

4. Now classify the plots based on their environmental characteristics, using the UPGMA algorithm. Please note that some of the environmental variables in the `dune.env` dataset are either ordinal or nominal (=factor). Which dissimilarity measure should we use to classify sites based on their environmental conditions? Build the dendrogram and compare it to the UPGMA dendrogram based on species data. How well do they match? [you may also compare the correlation of the cophenetic matrices]

```
#hints
?kmeans
?tapply
```

```
?hclust
?rect.hclust
?cophenetic
?cutree
?cluster::agnes
?FD::gowdis
```

PART 5 - UNCONSTRAINED ORDINATION

Q6 - Make yourself familiar with computing a pca in vegan

1. Run a PCA on the `varechem` dataset, both with and without standardizing the data. Examine the outputs and create Scree Plots. How much variation is explained by the first axis? [Advanced] How many axes are needed to retain 90% of total variation?

2. Create biplots for the PCA on transformed data, showing axes 1&2, and axes 1&3. Compare the two possible scalings. 3. Now run a NMDS on using the species data from`varespec`. Try to set k=2, and k=3, try with 20 iterations in both cases. Make sure you explore your output (what kind of object is it? where is the stress values stored?) How much does the stress get reduced by increasing the number of axes? [Make sure you start from a dissimilarity matrix appropriate to your data type!]

4. You can passively project your environmental predictors on your ordinations using the function `envfit`. Try to project on your 2-dimensional nmds the data contained in `varechem`. Explore your output. Which variable has the highest correlation with the first nmds axis? Try to plot your envfit onto your nmds.

5. [Advanced - optional] Re-run a k-means classification with 5 groups. Display these groups graphically when plotting your PCA ordination. Does the grouping make sense in the PCA?

```
#Hints
?rda
?screeplot
str(output.rda) #replace 'output.rda' with the name of your rda object
#You can extract the eigenvalues as
output.rda$CA$eig
?cumsum
?plot.cca
?biplot
?metaMDS
?envfit
```

PART 6 - CONSTRAINED ORDINATION

Also called "Direct gradient analysis" (in CANOCO) or "Canonical ordination" (in other literature). Remember that here we ask for the variation explained by constrains (predictors). We are using the same functions (cca, rda) but in the context of model formulas of the type $y \sim x + z$

Q7 - Now fit a constrained RDA to the varespec data, using a selection of three environmental variables as predictors. You may want to select your three variables after checking for multiple correlations for instance through the envfit plotting

1. As above, produce a screeplot, and explore the eigenvalues. What percentage of variation is explained by the predictors? Try to plot the output. Compare with the biplots produced for the unconstrained RDA.

2. Do the same on the `dune` data set. Use all environmental variables as predictors. How much variation do your predictors now explain? [hint - to include all variables use the formula: x ~.]

3. Plot the output. Note how nominal and ordinal variables [=factors] are shown.

4. Go back to your varespec model with three predictors, and compute the variation partitioning. Create a Venn diagram. How much variation is independently explained by your three predictors?

5. [optional] Repeat the constrained ordination analysis of `varespec` using CCA. How much variation is the

model now explaining?

```
#Hints
?varpart
?rda
?cca
?plot.varpart
```

You made it!