Table of Contents

Section: Ordination analysis

Explained variation and Monte Carlo permutation test (constrained ordination)

Theory R functions **Examples** Exercise 🔒

Example 1: Is the explained variance high enough to be considered as interesting?

This is direct continuation of Example 1 in tb-RDA section. We used Vltava river valley dataset to calculate tb-RDA (on log-transformed and then Hellinger transformed species composition data) with two measured variables (pH and SOILDPT) as explanatory. They explained 8.9% of the variance, and we may ask: if variables explain this amount of variance (which may seem rather low, less than 10%), is it enough to report this result?

To decide, we need to do two more things. One is to test whether the analysis is significant, meaning that this amount of variance is high enough compared to the variance generated (in average) by two random variables (this is the topic for Example 2 below). The other thing is to compare the value of explained variation (R2 = 8.9%) with the variation which would be explained by two best variables, i.e. variables which represent the strongest gradient along which the species composition is changing. As you can see in the theoretical part, we can create such variable for our dataset by applying first unconstrained ordination on our data and extract the first few unconstrained ordination axes (i.e. sample scores along them). These axes represent the main directions along which species composition changes the most rapidly, which is exactly what we need. Then we can use these axes as explanatory in the constrained version of the same ordination (here tb-RDA) to see how much variance they can explain. The number of axes I should take depends on the number of real variables I want to compare the variance with (in this case two).

Let's get data in and calculate tb-RDA on two explanatory variables (this is essentially repeating steps in Example 1 from tb-RDA:

```
vltava.spe <- read.delim
('https://raw.githubusercontent.com/zdealveindy/anadat-r/master/data/vltava-
spe.txt', row.names = 1)
vltava.env <- read.delim
('https://raw.githubusercontent.com/zdealveindy/anadat-r/master/data/vltava-
env.txt')
spe <- vltava.spe
env <- vltava.env[, c('pH', 'SOILDPT')]
library (vegan)
spe.log <- log1p (spe)  # species data are in percentage scale which is
strongly rightskewed, better to transform them
spe.hell <- decostand (spe.log, 'hell')  # we are planning to do tb-RDA,
this is Hellinger pre-transformation
tbRDA <- rda (spe.hell ~ pH + SOILDPT, data = env)  # calculate tb-RDA with
two explanatory variables
```

I can use the function RsquareAd from vegan to extract the explained variance directly from the ordination object:

```
R2.obs <- RsquareAdj (tbRDA)$r.squared
R2.obs
```

```
[1] 0.08868879
```

Result is 0.089 = 8.9%. Note that the function RsquareAdj returns a list with two components, r.squared and adj.r.squared, the first referring to the R2 we are interested now, and the second to adjusted R2, i.e. R2 corrected for the number of samples and number of explanatory variables (we will use this later).

Now we need to calculate the unconstrained version of tb-RDA on the same data and extract the first two ordination axes. The unconstrained version of tb-RDA is tb-PCA (i.e. PCA calculated on pre-transform data, where the pre-transformation should be the same in both analyses):

```
tbPCA <- rda (spe.hell)
tbPCA
```

```
Call: rda(X = spe.hell)

              Inertia Rank
Total          0.7048
Unconstrained  0.7048   96
Inertia is variance

Eigenvalues for unconstrained axes:
    PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8
0.09197 0.06075 0.04684 0.03537 0.02650 0.02361 0.02093 0.02035
(Showed only 8 of all 96 unconstrained eigenvalues)
```

For comparison, let's also include the summary of tb-RDA method:

```
tbRDA
```

```
Call: rda(formula = spe.hell ~ pH + SOILDPT, data = env)

              Inertia Proportion Rank
Total         0.70476    1.00000
Constrained   0.06250    0.08869    2
Unconstrained 0.64226    0.91131   94
Inertia is variance

Eigenvalues for constrained axes:
   RDA1    RDA2
0.04023 0.02227

Eigenvalues for unconstrained axes:
    PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8
0.07321 0.04857 0.04074 0.03144 0.02604 0.02152 0.01917 0.01715
```

```
(Showed only 8 of all 94 unconstrained eigenvalues)
```

You can see that total inertia (variance of the whole dataset) is identical in both analyses (0.70476). This is the variance we aim to explain.

Extract the two PCA axes from `tbPCA` object and use them as explanatory variables in tbRDA on the same data to see how much variance they explain:

```
PCA12 <- scores (tbPCA, display = 'sites', choices = 1:2)
tbRDA_PCA12 <- rda (spe.hell ~ PCA12)
RsquareAdj (tbRDA_PCA12)$r.squared
```

```
[1] 0.2167075
```

Two first tb-PCA axes, if used as explanatory in the tb-RDA on the same dataset, explain 21.7% of the variance. Note that this is the same number we would get by simply checking the amount of variance represented by the first two ordination axes in tb-PCA in the summary above: eigenvalue of PCA1 = 0.09197, eigenvalue of PCA1 = 0.06075, total variance (inertia) = 0.70476, recalculated into percentage: (0.09197+0.06075)/0.70476 = 21.7%. In fact, we don't need to really to the whole procedure (extract tb-PCA axes and use them as explanatory in tb-RDA), we can simply check the variance of *n* axes in unconstrained ordination (*n* = number of explanatory variables) and compare it with real variance explained by environmental variables.

The comparison here is: 8.9% explained by real variables (pH and SOILDPT) vs 21.7% which would be explained by two best, not correlated variables (if we had them). We see that measured variables explain something over 40% of variation they could (8.9/21.7 = 0.41), which is not bad. Remember important difference: PCA axes are (from definition) not correlated, while our real variables often will be (as in this case: `cor.test (~ pH + SOILDPT, data = env)` shows that Pearson's correlation coefficient between pH and SOILDPT is $r = 0.273$, and this correlation is significant at $P = 0.0069$).

The next step is to test whether the variation explained by our variables is significant - this is a topic for Example 1 in the Permutation test section.

Example 2: Is the variance explained by environmental variables significant?

This example directly follows the Example 1 in RDA & tb-RDA section and Example 1 in Explained variance section, consider checking them first. We used Vltava river valley dataset, and two field-measured environmental variables, soil pH and soil depth (`pH` and `SOILDPT`), to explain variance in species composition. We found that they can explain 8.9% of overall variance; when compared to the variance which can be maximally explained by two explanatory variables (21.7% explained by two tb-PCA axes, see here) this sounds not bad (it is more than 40% of variance we can maximally explain in this dataset with two variables). But is the result significant? By significant, I mean: is the variance considerably higher than the variance explained (in average) by two random variables not related to species composition? This is the task for Monte Carlo permutation test (check Theory part to see how it works).

First, get the data and calculate tb-RDA on them (this is esentially repeating beginning of Example 1 in the section Explained variance):

```
vltava.spe <- read.delim
```

```
('https://raw.githubusercontent.com/zdealveindy/anadat-r/master/data/vltava-spe.txt', row.names = 1)
vltava.env <- read.delim ('https://raw.githubusercontent.com/zdealveindy/anadat-r/master/data/vltava-env.txt')
spe <- vltava.spe
env <- vltava.env[, c('pH', 'SOILDPT')]
library (vegan)
spe.log <- log1p (spe)
spe.hell <- decostand (spe.log, 'hell')
tbRDA <- rda (spe.hell ~ pH + SOILDPT, data = env)
R2.obs <- RsquareAdj (tbRDA)$r.squared
R2.obs
```

```
[1] 0.08868879
```

In the next step, calculate variance explained by randomized env. variables

```
env.rand <- env[sample (1:97),]  # the function "sample" will reshuffle the rows with environmental variabels
tbRDA.rand <- rda (spe.hell ~ pH + SOILDPT, data = env.rand)
RsquareAdj (tbRDA.rand)$r.squared
```

This value represents the variance explained by two random explanatory variables. My result was 0.01854349, but this value will change in each run. We need to do enough repetitions (permutations) to get an idea about the distribution of this values (null model). We can use the “for” loop for it, or, as in this case, function “replicate”, with two arguments: n = number of replicates, and `expr` = expression to be replicated (if more lines of script are involved, this expression needs to be enclosed in curly brackets {}):

```
n.perm <- 99  # set the number of permutations
R2.rand <- replicate (n = n.perm, expr = {
  env.rand <- env[sample (1:97),]
  tbRDA.rand <- rda (spe.hell ~ pH + SOILDPT, data = env.rand)
  RsquareAdj (tbRDA.rand)$r.squared
  })
<code>
The vector ''R2.rand'' contains 99 values of variance explained by random variables. In the next step, we will merge them with the observed R2 (''R2.obs''), since this
```