Common confusions and mistakes

A temporary list of common confusions and mistakes when applying numerical methods in community ecology. Based on fixing final reports of students in the class Numerical Methods in Community Ecology.

Choosing between transformation-based ordination methods vs linear x unimodal methods

Wrong: you first use DCA to decide whether to use linear or unimodal ordination methods, and if you decide for linear methods, you Hellinger-transform the data and call it transformation-based ordination.

This confuses two alternative approaches how to do ordination on community data - read further.

There are three alternative ways how to analyse community data.

1. Use either linear or unimodal method (PCA vs CA,DCA in case of unconstrained ordination, RDA vs CCA in case of contrained) to analyze the data; the decision between both is done based on heterogeneity of the species composition dataset, and this heterogeneity (beta diversity) can be measured by DCA (the length of the first DCA axis is a measure of heterogeneity; if it is longer than 4, the dataset is considered as heterogeneous and suitable for unimodal methods, if < 3 S.D., it can be considered as homogeneous, suitable for linear methods; 3-4 - in the middle, both linear or unimodal methods are ok). After this decision, use either linear (PCA, RDA) or unimodal (RDA, CCA) method to analyse data, do not use tb-PCA or tb-RDA.
2. Use transformation-based ordination method. You do not need to make a decision between linear or unimodal methods, simply you opt for transformation-based method. Transform the species composition data by Hellinger transformation (there are also other transformations, but this is the most often used), and analyse data either by unconstrained (tb-PCA) or constrained (tb-RDA) approach, depending on your purpose. No DCA is done before, it is not necessary, since you are not making decision between linear or unimodal ordination method (you use linear method applied on Hellinger transformed data).
3. Use distance-based methods (PCoA and NMDS for unconstrained, and db-RDA for constrained ordination).

Hellinger transformation has therefore two types of use: (a) to calculate transformation-based ordination, or (b) to standardize species composition data:

- Ad (a): linear ordination methods (PCA, RDA) are based on Euclidean distances, which are sensitive to double-zero problem. If the species composition data are first Hellinger transformed, and then used in linear ordination methods, the combination of Hellinger transformation + Euclidean distance means that the distance used by these methods is Hellinger distance, which is not sensitive to double-zero problem. Here, the reason to transform the data is to simply avoid influence of double-zeros in the ordination analysis.
- Ad b) Hellinger transformation converts species abundances from absolute to relative values (i.e. standardizes the abundances to sample totals) and then square roots them. This could be useful if we are not interested in changes of absolute species abundances, but relative abundances. E.g. if species A has abundance 30 in sample 1 and 20 in sample 2, and species B

has abundance 3 in sample 1 and 2 in sample 2, the absolute change is 10 for species A and 1 for species B, but relative change is the same for both. If this is the purpose of this transformation, then Hellinger transformed data could be used also in other ordination methods, e.g. CA or CCA; then, it is already not “transformation-based ordination” in the sense of point (a) above.

Confusing "analysis of species richness" and "analysis of species composition"

Wrong: you analyse species composition (e.g. using ordination or cluster analysis), but you talk about analysis of species richness.

Analyzing pattern of species richness (or diversity) and species composition are two different things, although it could be sometimes interesting to analyse both in parallel. If we analyse species richness, we simply ask which community is more and which less diverse, and which factors are the best to interpret this pattern. In contrast, if we analyse pattern in species composition, we ask how does species composition change (in space or time) and which factors are best to explain these changes. Changes in species composition may not necessarily be accompanied by changes in species richness (two communities may have the same number of species, i.e. the same richness, yet completely different species composition).

There are many general patterns of species richness, e.g. along latitude (diversity decreases from Equator toward poles), along productivity (often reported as hump-back pattern with highest diversity in intermediate levels of productivity) or along pH (for certain organisms, e.g. plants or mollusks, pH is important factor influencing richness - basic habitats are more species rich than acid, although this works only in some geographic regions, like Europe and North America). Often the same factors influence also changes in species composition, but mechanisms behind are likely different.

Use “envfit” to project explanatory variables into RDA or CCA and to test their significance

Wrong: you calculate constrained ordination (e.g. RDA, CCA), and then you use the function ‘’envfit’’ to project explanatory variables onto the ordination diagram (theoretically it can be done, but needs to be done in a special way, see below), or even test the significance of these variables.

Simple suggestion: The function `envfit` is designed to calculate regression of “supplementary” (not “explanatory”) variables on ordination axes of unconstrained ordination, and test the significance of this regression by permutation test. This function is not suitable to project “explanatory” variables onto ordination diagram of constrained ordination (although theoretically it can be done, see below) and definitely not to test the significance of individual variables in this way (see example below showing that in that case all variables will usually be highly significant, because constrained ordination axes already contains their information).

Longer explanation follows below, with examples.

First, to clarify. One of the use of the function ‘’envfit’’ (library vegan) is to calculate regression of supplementary variables to ordination axes in unconstrained ordination, in order to help with

interpretation of these axes. Say we have community data and three external environmental variables; we calculate unconstrained ordination using the community data, and then we aim to interpret the ecological meaning of unconstrained ordination axes with available environmental variables by calculating their regression. These variables can then be projected onto the ordination diagram as supplementary (not explanatory!), and regression of each variable independently can be tested by permutation test (this is what ''envfit'' returns). But variables which are significant here are not those which are important for species composition, but those which have the best fit to variation extracted by unconstrained ordination into the main (usually the first and the second) ordination axes. Often these variables are important, but not necessarily.

Now, if you calculate constrained ordination, then environmental variables enter the analysis, and resulting (constrained) axes already contain information about them. If now you use ''envfit'' to calculate regression of these same variables to constrained axes, you are likely to get highly significant results; you are regressing two variables (environmental variable vs samples scores on ordination axes) which contain the same information (since axes were calculated also using these environmental variables).

Theoretically, `envfit` can be used to project arrows or centroids with explanatory variables onto diagram of constrained ordination, but the fitting must be done on linear scores of ordination axes, not on sample scores, otherwise the arrows will have different direction then they should be (see example below).

Examples:

It can be shown that arrows for environmental variables projected by ''envfit'' are different from the true direction of these variables in ordination diagram, unless they are calculated properly (on linear combination of scores). In the diagram below (based on river valley data), the three variables (pH, SOILDPT and ELEVATION) have been used as explanatory variables in tb-RDA. The left diagram shows the true direction of explanatory variables. The right diagram shows three options of using `envfit` function. Blue arrows show the true direction (`envfit` calculated with argument `display = 'lc'` , i.e. on linear combination of scores). Red arrows indicate direction if they are added using ''envfit'' fitted to samples scores (by default only the first and second ordination axis), and green arrows show the same with ''envfit'' relating them to three instead of two axes.

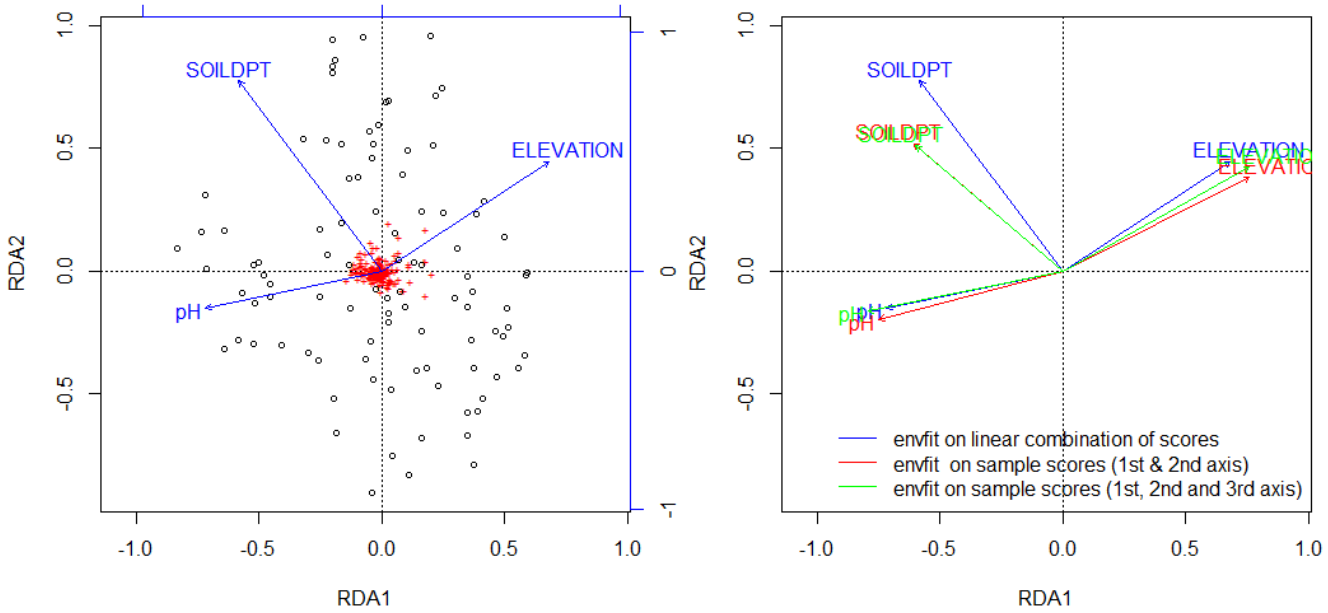
```
vltava.spe <- read.delim
('https://raw.githubusercontent.com/zdealveindy/anadat-r/master/data/vltava-
spe.txt', row.names = 1)
vltava.env <- read.delim
('https://raw.githubusercontent.com/zdealveindy/anadat-r/master/data/vltava-
env.txt')

library (vegan)

spe.hell <- decostand (log1p (vltava.spe), method = 'hellinger')
tbRDA <- rda (spe.hell ~ pH + SOILDPT + ELEVATION, data = vltava.env[,1:18])

par (mfrow = c(1,2))
ordiplot (tbRDA)
ef <- envfit (tbRDA ~ pH + SOILDPT + ELEVATION, data = vltava.env, display =
'lc')
ef12 <- envfit (tbRDA ~ pH + SOILDPT + ELEVATION, data = vltava.env)
ef123 <- envfit (tbRDA ~ pH + SOILDPT + ELEVATION, data = vltava.env,
```

```
choices = 1:3)
ordiplot (tbRDA, type = 'n')
plot (ef, col = 'blue')
plot (ef12, col = 'red')
plot (ef123, col = 'green')
legend ('bottomright', lwd = 1, col = c('blue', 'red', 'green'), legend =
c('envfit on linear combination of scores', 'envfit  on sample scores (1st &
2nd axis)', 'envfit on sample scores (1st, 2nd and 3rd axis)'), bty = 'n')
```



Now about testing the significance (to show why it should not be done using the function `envfit`). In the example below, I again used the river valley dataset. I generated one random variable, i.e. I assign a random value to each site, and use it as “environmental variable” (obviously without any ecological meaning and without any relationship to the species composition). I calculated tb-PCA, and used envfit to fit the random variable to ordination axes (result not significant, as expected); then I calculated tb-RDA with this random variables as explanatory and tested significance of this RDA (not significant, also as expected). Then I use ‘’envfit’’ to fit the random variable to ordination axes (the result is highly significant, which could indicate that the random variable is important, although it is not). The result shows that to use `envfit` to fit environmental variables to RDA which is based on the same environmental variable is highly misleading.

```
spe.hell <- decostand (log1p (vltava.spe), 'hell')
set.seed (123)
random.var <- runif (97)  # generate single random variable
tbPCA <- rda (spe.hell)
envfit (tbPCA ~ random.var)  # not significant
# ***VECTORS
#
# PC1     PC2     r2 Pr(>r)
# random.var 0.41445 0.91007 0.0072  0.716
# Permutation: free
# Number of permutations: 999
```

```
tbRDA.rand <- rda (spe.hell ~ random.var)
anova (tbRDA.rand)  # the RDA is not significant (since the explanatory
variable is random and has no meaning)
# Permutation test for rda under reduced model
# Permutation: free
# Number of permutations: 999
#
# Model: rda(formula = spe.hell ~ random.var)
# Df Variance      F Pr(>F)
# Model     1  0.00724 0.9861  0.472
# Residual 95  0.69752

envfit (tbRDA.rand ~ random.var)  # the fit of random variables to first two
ordination axes is significant, since the first axis is calculated from it
# ***VECTORS
#
# RDA1      PC1     r2 Pr(>r)
# random.var  0.92615 -0.37716 0.4525  0.001 ***
#   ---
#   Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
# Permutation: free
# Number of permutations: 999
```

Check for missing values in environmental variables before conducting constrained ordination (RDA, CCA)

Problem: if the environmental variables contain missing values and such matrix is used in constrained ordination as explanatory, the samples containing (even just one) missing values will be removed from analysis. If you have many environmental variables and each has some missing values (in the worst case each variable missing for different samples), such data with many holes may mean that the final analysis is based on rather reduced number of samples.

Solution: function `na.omit` applied on environmental matrix will remove all samples with one or more missing values. The same samples need to be removed also from species data. Make sure that you know on how many samples your analysis is based! Also, avoid replacing missing values by zero; this could make sense in case of species data (missing species has zero abundance), but usually does not make sense in case of environmental variables (if soil depth was not measured and has missing values, replacing it by zero would mean that the vegetation grows on a rock, which is not true; the same for e.g. soil chemical concentrations - not measured does not mean not present).

Doing constrained ordination analysis with all explanatory variables, but using only those selected by forward selection to display in ordination diagram

Mistake is if you calculate constrained ordination (RDA, CCA, tb-RDA) with all explanatory variables, then do forward selection to select only those significant, and then you create ordination diagram

based on analysis with all explanatory variables, but in this ordination diagram you display only those explanatory variables which were significant in forward selection.

Correct way would be to display ordination diagram which is based on constrained ordination using only the selected (by forward selection) environmental variables as explanatory variables. Indeed, the ordination diagram based on all environmental variables, and the ordination diagram based on only a small subset of selected variables will look different, and different number of ordination axes will be constrained (since number of constrained axes = number of env. variables, if they are quantitative). You should, however, still conduct constrained ordination with all environmental variables before you do forward selection (so called "global test", and only if it is significant, you can proceed to forward selection itself); there is, however, perhaps no need to draw ordination diagram with all explanatory variables (unless there is a clear interpretational need to do that).

Significance of constrained ordination is not tested by ANOVA

Problem: statement in the Methods that "the significance of the constrained ordination was tested by ANOVA".

Why: Variance explained by constrained ordination is tested by a permutation test, in which the real explained variation is compared with the variation explained by randomised explanatory variables. This kind of test is generally called Monte Carlo permutation test (thus, Monte Carlo test is not only specific for constrained ordination, the same family of tests could be used to test e.g. correlation of two variables). If the observed explained variation is higher than some proportion (usually 95%) of variation explained by randomised variables, it is considered as high enough to be significant (at $P < 0.05$). It gets a bit more complicated; the test statistic, in fact, is not the explained variation (R^2), but so-called pseudo-*F* value, which consists of explained variation and number of degrees of freedom (which reflects number of explanatory variables as well as covariables). The test then resembles ANOVA in a way that ANOVA (analysis of variance) is also based on F-value. In `vegan`, the function conducting Monte Carlo permutation test is called `anova`, for mostly historical reason; the description of the function (see `?anova.cca`) states that it is "ANOVA-like permutation test for Constrained Correspondence Analysis (cca), Redundancy Analysis (rda) or distance-based Redundancy Analysis (dbRDA)". The term ANOVA is traditionally understand as an analysis of variance with a single dependent variable, so nothing related to constrained ordination.

Conclusion: You can mention that the constrained ordination was tested by permutation test, Monte Carlo permutation test, or ANOVA-like permutation test (the last one is perhaps the least common). But do not say simply that the significance of constrained ordination was tested by ANOVA, this would be misleading.

In cluster analysis, Ward method cannot be combined with Bray-Curtis distances (unless square-rooted)

Ward's algorithm is based on measuring the distance of the sample to the centroid of the group of samples in orthogonal Euclidean space. This means that distances which are not Euclidean should not be directly used with this algorithm unless modified ("being Euclidean" here means that the distance obeys triangular inequality principle and can be displayed in orthogonal Euclidean space; e.g.

Manhattan distance is not Euclidean, but the square root of Manhattan distance is Euclidean). Strictly speaking, Ward's clustering algorithm should be used only with Euclidean distances, and since Bray-Curtis distance is not Euclidean, it should square-rooted first before used with Ward's algorithm.

Using the first unconstrained axis in constrained ordination to see how much variation can be maximally explained by a single explanatory variable

Problem: You want to know how much variation you can maximally explain by a single explanatory variable in constrained ordination (RDA, tb-RDA, CCA), and for this, you will check the variation represented by the first unconstrained ordination axis of this constrained ordination (with the environmental variable used as explanatory).

Explanation: If you want to know how much variation you can maximally explain by a single explanatory variable, you follow this logic: the ordination axis of unconstrained ordination (e.g. PCA1) represents the direction of the strongest compositional gradient, and can be (theoretically) considered as a perfect explanatory variable for the dataset from which it was calculated. You can use it as explanatory in the constrained variant of the same ordination method (e.g. RDA) to see how much variation it explains; the same value you get if you simply check the variation represented by this axis in the original unconstrained ordination (the eigenvalue of the axis divided by total inertia/variance, Fig. 1, the horizontal axis in the right panel). This, however, does not apply to the first unconstrained ordination axis of the constrained ordination (e.g. the PC1 in RDA, Fig. 1, the vertical axis in the left panel), in which some environmental variable (or variables) was used as explanatory. Such axis represents the maximum variance a single explanatory variable can explain in the dataset after the variation of the explanatory variable has been removed.

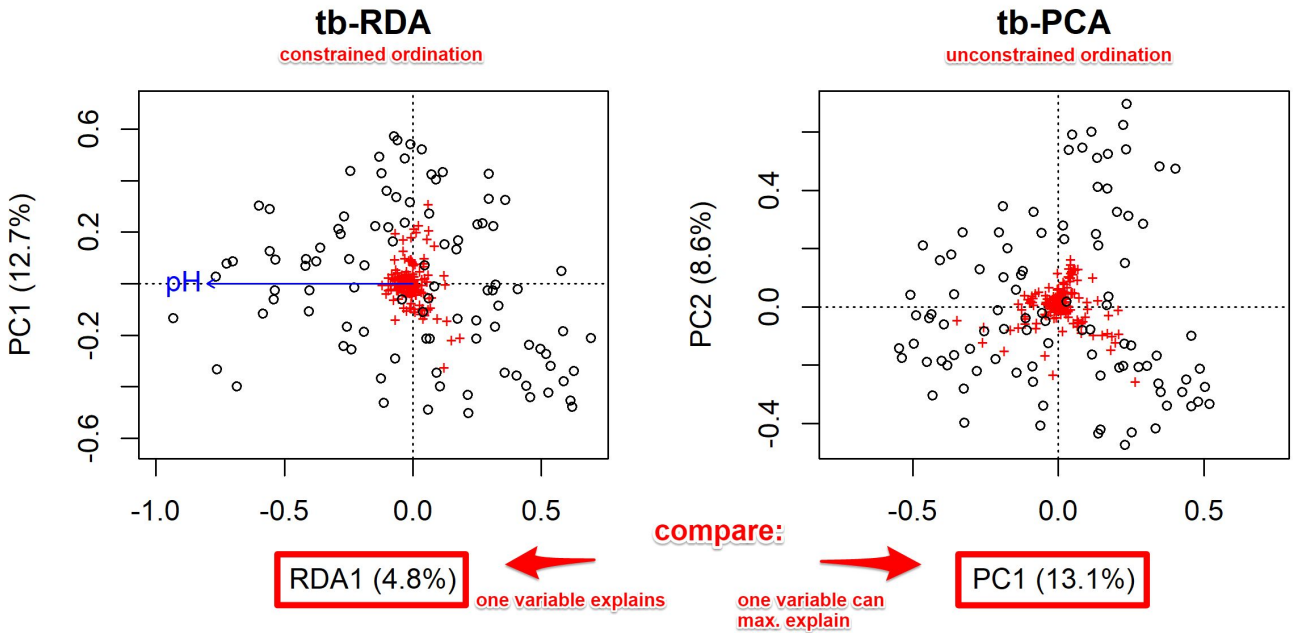


Figure 1