Section: Ordination analysis

Explained variation and Monte Carlo permutation test (constrained ordination)

Theory R functions Examples

Constrained ordination includes multivariate regression analysis, and as in ordinary least squared regression, the effect size is measured by R^2, the coefficient of determination. R^2 quantifies the variation in species composition explained by a linear model of the environmental variable(s), and can be calculated (if no covariables are included) from the analysis output as the sum of eigenvalues of constrained axes divided by the total variation (sum of all eigenvalues). The value of R^2 in constrained ordination suffers from the same drawback as in ordinary regression, namely that it increases with the number of explanatory variables and decreases with the number of samples in the dataset, making the values incomparable between analyses done on different datasets. The solution is to **use adjusted R^2**. The value of explained variation itself is not too informative unless it is compared with the variation the same number of explanatory variables *could possibly* explain (which is usually far from 100%). Also, even if the explanatory variables are in fact randomly generated, the R^2 is non-zero and positive (in contrast to adjusted R^2, which may be zero or even negative), and to decide whether the results are interpretable, it is useful to test their **significance by Monte Carlo permutation test**.

Adjusted R2

R^2 is known to depend on the number of samples in the dataset (sites in our case) and on the number of explanatory variables: with the number of samples R^2 decreases, and with the number of predictors (even if these are randomly generated) it increases (Fig. 1). The relationship can be expressed numerically: *p* random predictors explain (in average) *p*/(*n*-1) of the variation (where *n* is the number of samples in the analysis).

The solution to this problem is to calculate adjusted R^2. For linear ordination methods (as well for ordinary least squared multiple regression) the adjusted R^2 can be calculated using Ezekiel's formula:

$$R^2_{adj} = 1 - \frac{n-1}{n-p-1}\left(1-R^2\right)$$

where *n* is the number of samples and *p* is the number of predictors (explanatory variables). Resulting adjusted R^2 are independent on the number of samples and predictors (Fig. 1).

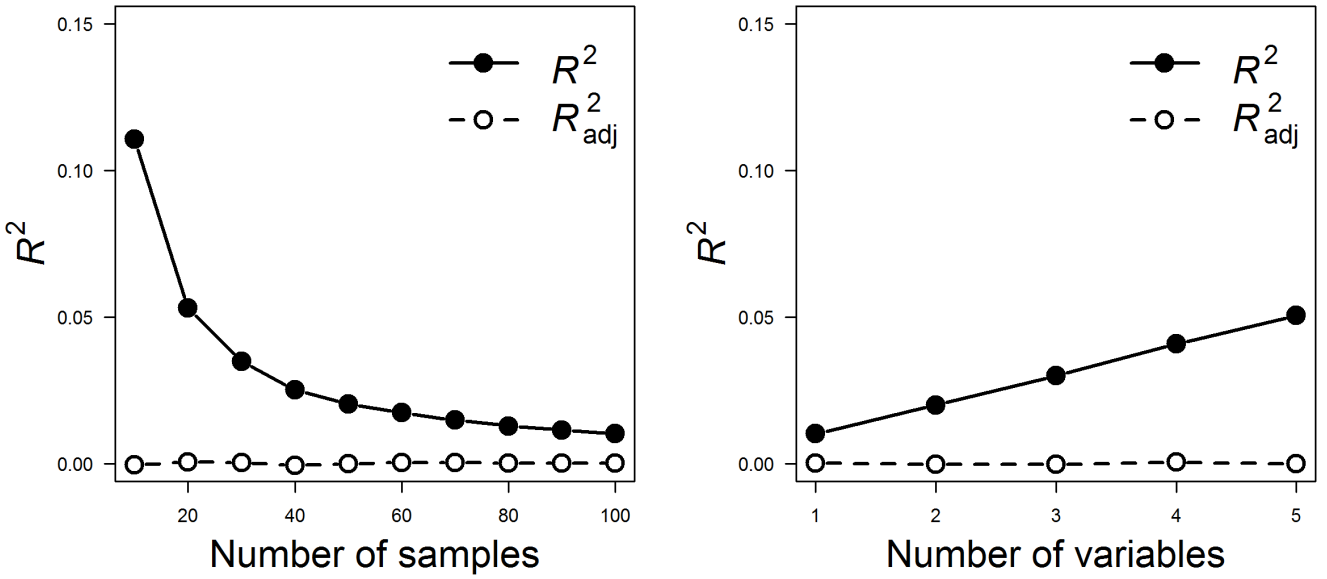


Figure 1: Comparison of variance explained in constrained ordination expressed by R2 and adjusted R2. The community data with one strong gradient were simulated using the library (simcom), with an increasing number of samples. The explanatory variables are randomly generated. R2 decreases with the number of samples in the dataset (left figure) and increases with the number of explanatory variables (although these are just randomly generated). Adjusted R2 is not influenced by these two dataset parameters.

In the case of unimodal ordination methods, however, the values returned by Ezekiel's formula are overestimated (and the dependence of variation on the number of samples and/or explanatory variables is not removed), and the R^2 needs to be adjusted using the permutational method proposed by Peres-Neto et al. (2006). The principle of this permutation adjustment is based on using modified Ezekiel's formula to compare observed variation explained by the variables (R^2) with expected (mean) variation the same number of variables would explain if they are random (R^2_{perm}, Fig. 2). Adjusted R^2 calculated by the permutation method will slightly differ among calculations (these differences will be rather small if the number of permutations is set to be high).

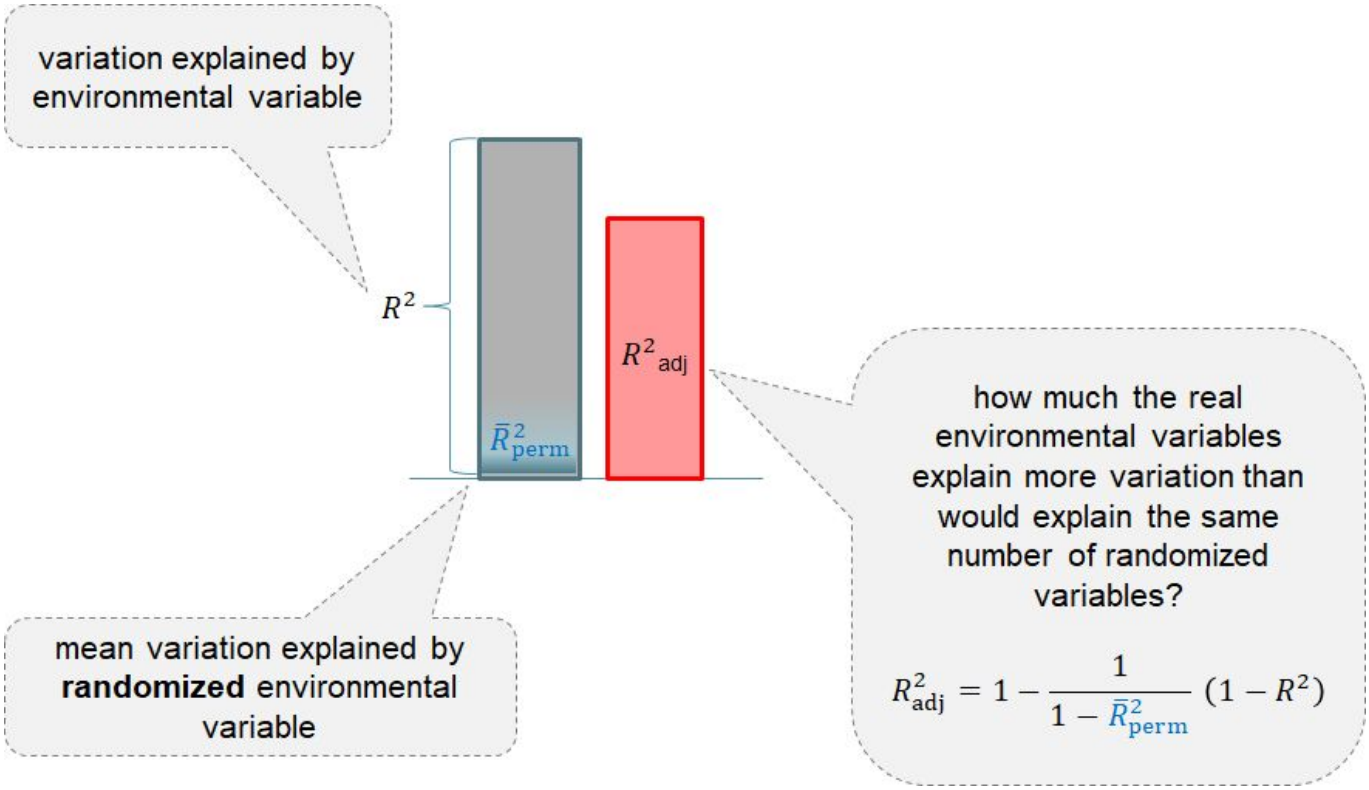


Figure 2

In contrast to R^2, the values of adjusted R^2 can reach zero, which means that the explanatory variables do not explain any variation in species composition, or they can be even negative, which means that explanatory variables explain even less variation than would be explained (in average) by the same number of randomly generated ones. The negative values are usually ignored and not interpreted (this is important e.g. when interpreting fractions in variation partitioning).

Is the value of explained variation too low?

The variation explained by constrained ordination may often seem as too low in the absolute terms. For example, in Example 1 in this section, the variation in species composition of vltava dataset explained by two explanatory variables, soil pH and soil depth, is less than 9%. Are the results of the analysis explaining less than 9% of variation worth to interpret/think of/publish?

To correctly interpret the value of explained variation, you need to consider that in the case of multivariate linear regression (which constrained ordination is), two explanatory variables will be unable to explain 100% of the overall variation. In fact, the amount of variation explainable by a given number of explanatory variables in the case of the certain dataset can be exactly calculated, if we assume that the best explanatory variable for that dataset is represented by sample scores on the ordination axes calculated by an unconstrained variant of the ordination on the same dataset. For example (Fig. 3), if we take one explanatory variable, and we want to know how much it could maximally explain in the constrained ordination analysis done on given dataset (soil pH used as explanatory in tb-RDA on log and Hellinger transformed species composition data from vltava dataset, explaining 4.8% of variation), we can do the following: 1) calculate unconstrained variant of given ordination method (here tb-PCA) on the same species composition dataset (with the same transformation of raw data if applicable), and 2) check the variation represented by the first ordination axis (calculated by dividing the eigenvalue of this axis by total variation in the dataset). This value (13.1% in our example) is the maximum variation any one predictor can explain in a given dataset. In our example, pH explained 4.8% of total variation, while it could maximally explain 13.1%

- in fact, it explained more than third of what it could explain (4.8/13.1= 36.6%). This is not too bad.

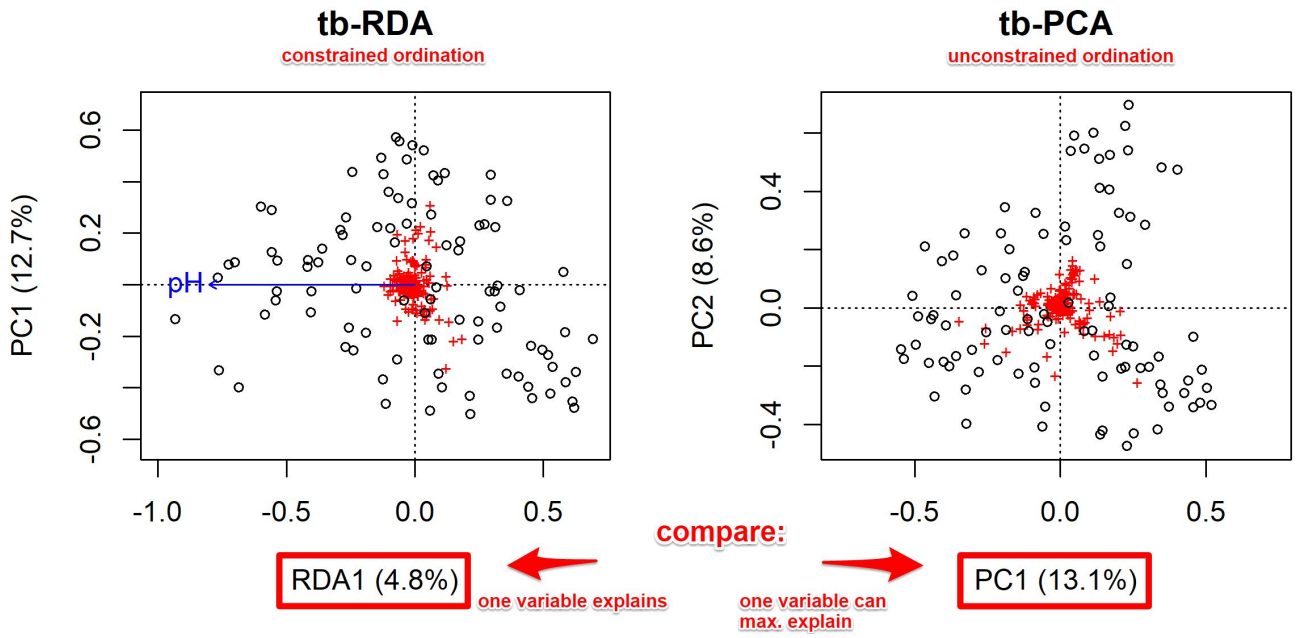


Figure 3

The same applies for more than one explanatory variable – e.g. if you have two variables, you take the variation represented by the first two axes in the unconstrained variant of the ordination for comparison. In the case of more than one variables, however, the comparison starts to have the problem that ordination axes are by definition not correlated, while explanatory variables are (often) somewhat correlated. It means that the variation represented by *n* unconstrained ordination axes is the maximum variation which could be explained by *n* explanatory variables which are *not intercorrelated*.

Monte Carlo permutation test

Variation explained by explanatory variables in constrained ordination is routinely tested and often only variables (or sets of variables) which are significant are being considered for interpretation. The test is based on comparing observed variation explained by environmental variables with variation explained by (the same number of) randomized variables. The test statistic used is the so-called **pseudo-F value**, which is calculated from the axis eigenvalue, the residual sum of squares of the constrained model, numbers of explanatory variables, covariables and samples in the dataset. If not covariables are used in the analysis, pseudo-F value can be replaced by the values of raw R^2.

Monte Carlo permutation test, in general, is a permutation test which compares the observed value of the test statistic with the expected distribution of the test statistic generated by permuting the original data under the assumption of the null hypothesis. In the case of constrained ordination, the test statistic is pseudo-F (but can be also R^2 if no covariables are used); the observed value of the test statistic (F_{data}) is from the analysis using the original explanatory variables, and the distribution of the null expectations for the test statistic (F_{perm}) are values from the analysis using the permuted explanatory variables (permutation removes the link of explanatory variables to species composition data, creating assumptions of the null hypothesis about not relationship between explanatory variables and species composition). The permutation needs to be repeated many times to build the null distribution to which the observed value of the test statistic is compared (Fig. 4). The resulting P-

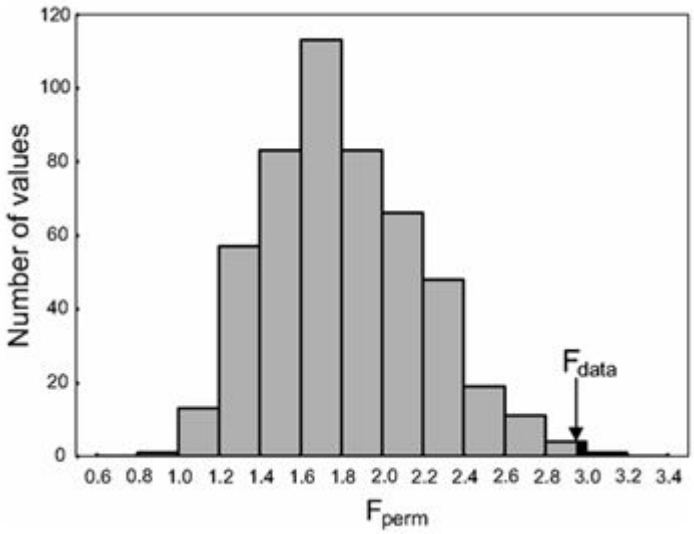
value is calculated as

$$P = \frac{n_x + 1}{N + 1},$$

where n_x is the number of permutations where the expected value of the test statistic is higher than observed value ($F_{perm} > F_{data}$), and N is the number of all permutations. The formula also explains why the numbers of permutations are usually set to end with 9 (e.g. 499 instead of 500): the observed value of the test statistic is considered to be also the part of the null distribution and is therefore added together with expected values (and the number one is added to the formula).

The lowest P-value you can reach depends on the number of all permutations (N) and can be calculated as $P_{min} = 1/(N+1)$. For example, with N = 199 permutations, the lowest P-value which can be obtained is 1/(199+1) = 0.005, and one cannot, therefore, hope to reject the null hypothesis at P < 0.001. Especially if the multiple testing correction is applied (e.g. in the case of forward selection or testing several constrained axes), it may be necessary to increase the number of permutations into a quite high value to make sure that the minimum P-value after adjustment will be lower than the selected significance threshold (example: with 499 permutations, $P_{min} = 1/(499-1) = 0.002$; if we are conducting 10 tests (e.g. because we do forward selection from 10 explanatory variables in constrained ordination) and we apply Bonferroni correction, each P-value gets multiplied by 10, and the minimum achievable P-value after adjustment is $P_{min-adj} = 1/(499-1)*n\_tests = 0.02$; we may need to increase the number of permutations to be able to reject the null hypothesis at lower P-values).

Figure 4: Observed value of the pseudo-F value (Fdata) compared with the null distribution of expected values (Fperm). From Šmilauer & Lepš (2014).



Several alternative tests are available. One can test the **significance of only the first constrained axis** or **all constrained ordination axes** (in the case of only one explanatory variable, a test of the first and all constrained ordination axes is identical). It is also possible to test **each constrained ordination axis** separately, e.g. to decide how many axes should be displayed in the ordination diagram. If each axis is tested separately, consider correcting resulting P-values for the multiple testing issue (e.g. by Holm correction). Unconstrained ordination axes cannot be tested.