Variable selection (constrained ordination)

Theory R functions Examples

Variable selection is a procedure for selecting a subset of explanatory variables from the set of all variables available for constrained ordination (RDA, CCA, db-RDA). The goal is to reduce the number of explanatory variables entering the analysis while keeping the variation explained by them to the maximum. Variable selection is suitable mostly in case of observational studies, where many (often highly intercorrelated) environmental variables are recorded, to reduce their number (and to simplify the story); it is usually not useful for experimental studies with the balanced design of treatment application.

The standard method is *forward* selection, which is adding explanatory variables one by one; *backward* selection, in contrary, starts from the full model (with all variables) and deletes variables which the least decreases the total explained variation. Combination of both approaches is *stepwise* (forward-backward) selection, in which in every step the analysis checks whether some of the already included variables cannot be removed to improve the model.

The simplified sequence of steps in the case of *forward selection* is the following:

1. first, test the significance of the global test with all explanatory variables included; if it is significant, you may proceed to forward selection, while if it is not, it is better not to (even with randomly generated explanatory variables you have rather good chance to select some of them as significant during forward selection).
2. use each variable one by one as explanatory in constrained ordination, and record the explained variation (this variation represents a simple (or marginal) effect of each variable);
3. sort variables according to variation explained by them with the highest values at the top;
4. check whether the variation explained by the best variable is significant using Monte Carlo permutation test - if yes, include it to the model, if not, stop the selection;
5. use each of remaining explanatory variables and check how much variation they (each separately) explain if put as explanatory (with the already selected variable acting as covariable);
6. sort again the variables according to the decreasing variation explained by them (now this variation represents the partial effect of this variable) and choose the one explaining the most; test whether the variation is significant, and if yes, select it into the model; if not, stop the selection;
7. continue by step 5 until the variation explained by the best variable is not significant.

The significance of the variables (step 7 above) is one of the possible stopping rules (once the best remaining variable is not significant, the selection is stopped). Alternative stopping rule is reaching the adjusted R^2 of the global model (Blanchet et al. 2008): first, calculate adjusted variation explained by all explanatory variables (global model); if during the forward selection the adjusted variation explained by selected variables reaches the R^2_{adj} of the global model (with some given precision threshold), the selection will be stopped (available in function `ordiR2step` in `library (vegan)` and `forward.sel` in `library (adespatial)`).

To reduce the risk of Type I error due to conducting a set of tests of significance during the forward selection, resulting P-values may need to be adjusted, e.g. by Holm or FDR correction, and only variables significant at given alpha should be considered.

In the Example 1 with the chemical variables measured in fen water and their relationship to the species composition of Carpathian wetlands, forward selection chose a subset of five variables (after adjusting the P-values with Holm's correction):

	variables	R^2	Cum R^2	Cum R^2_{adj}	F-value	P-value	P-value (Holm)
1	Ca	0.139	0.139	0.126	10.976	0.00002	**0.00028**
2	conduct	0.032	0.171	0.147	2.627	0.00004	**0.00052**
3	Si	0.027	0.199	0.162	2.243	0.00016	**0.00192**

	variables	R^2	Cum R^2	Cum R^2_{adj}	F-value	P-value	P-value (Holm)
4	NH3	0.024	0.223	0.175	2.006	0.00080	**0.00880**
5	NO3	0.021	0.244	0.185	1.787	0.00304	**0.03040**
6	Mg	0.019	0.263	0.193	1.637	0.00824	0.07416
7	pH	0.017	0.280	0.199	1.503	0.02094	0.16752

The table is a simplified output of the function `forward.sel` (or similarly also `ordiR2step`). It contains the variables in the order as they were selected during the forward selection; **R^2** is the partial variation the variables explains (i.e. variation the variable explains after accounting all previously selected variables as covariables); **Cum R^2** and **Cum R^2_{adj}** are cumulative variance (not-adjusted and adjusted R^2) explained by given variable together with all previously selected variables. **F-value** is the pseudo-F of given variable, and **P-value** is the original P-value which was used to decide whether the variable should be selected; **P-value (Holm)** is the P-value adjusted for multiple testing issue by Holm's correction (adjusting for all potential tests which may have been done, i.e. for the number of all variables from which the selection is done - 14 in this case; **P < 0.05 are displayed in bold**). The number of permutations in the test of significance was set very high (49,999), meaning that the lowest P-value is 1/(49,999+1) = 0.00002, and even after the Holm's correction, this P-value is 0.00028, i.e. well below 0.001. Results show that five variables (Ca, conductivity, Si, NH3 and NO3) were selected, together explaining R^2_{adj} = 18.5% (compared to 20.1% explained by global model with 14 variables). Comparison of tb-RDA ordination diagram of global model and model with forward-selected variables is on Fig. 1.

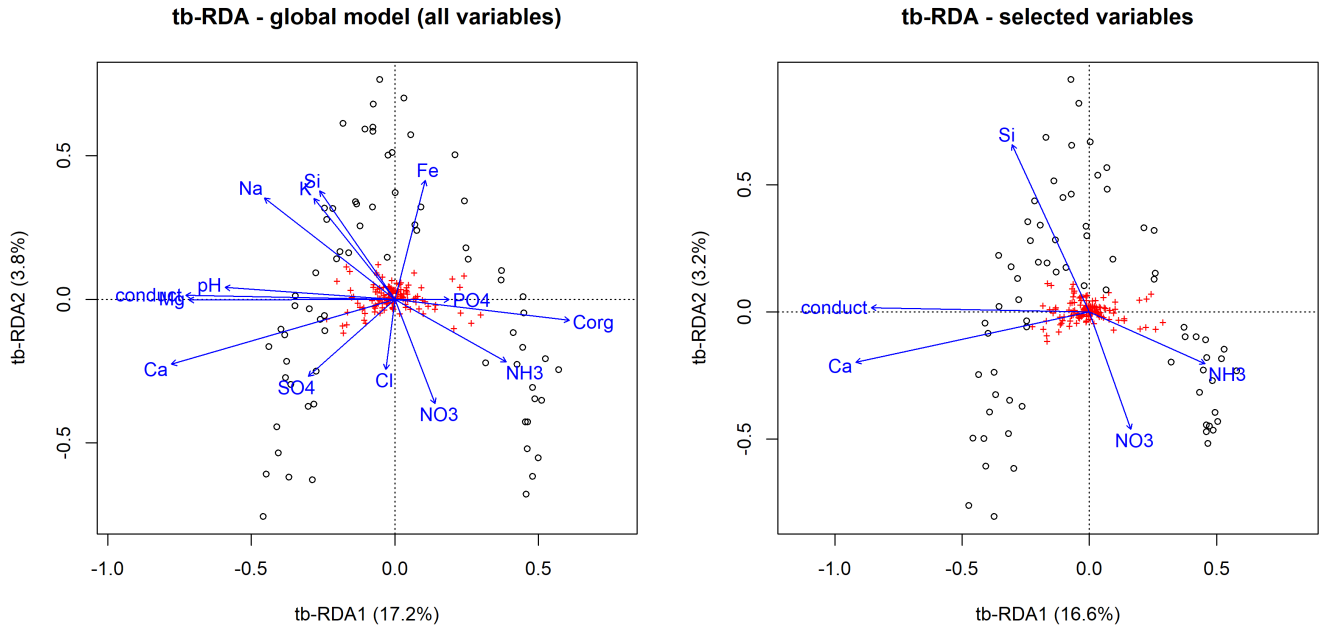


Figure 1

From:
https://anadat-r.davidzeleny.net/ - **Analysis of community ecology data in R**

Permanent link:
https://anadat-r.davidzeleny.net/doku.php/en:forward_sel

Last update: **2021/03/03 19:54**