Multivariate statistics in R

Hannes PETER
Martin BOUTROUX
Zhe LIU

Another important «discovery»!

- Discrepancy/error in «scaling» (e.g. PCA) between slides and R code!

- **Scaling 1**
  - focus on «sites/objects»
  - distances between objects represent euklidean distance (PCA)
  - angles between vectors should not be interpreted

- **Scaling 2**
  - focus on «species/variables»
  - distance between objects should not be considered to be euklidean distance (PCA)
  - angles between vectors reflect their correlation
    - angle of 90° between vectors reflects no correlation
    - angle of 20° reflects strong positive correlation

- in vegan: *biplot(pca, scaling="sites")*

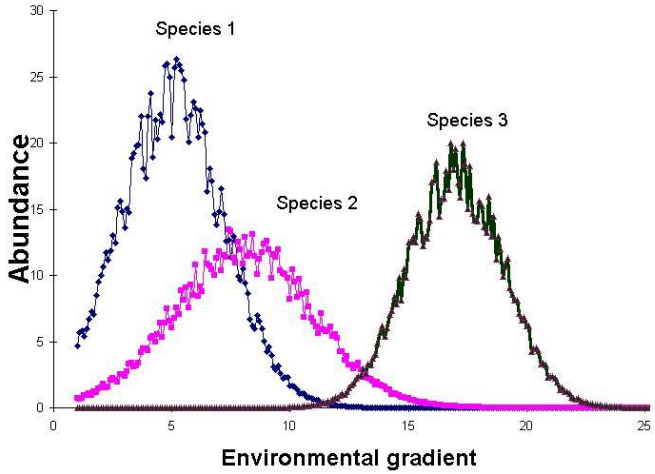
Outlook

- The horseshoe/arch effect
  - An artefact that arises in unconstrained ordination when there is complete turnover in an assemblage along a single environmental gradient
  - **D**etrended **C**orrespondence **A**nalysis removes arch effects

- Constrained ordination
  - analogous to supervised classification – use explanatory variables to explain variation in response variables
  - useful to test hypotheses or to detect trends that are «hidden» by high variability; ability to use «conditions» (partial RDA)
  - constrained ordination techniques:
    - Redundancy Analysis (RDA) (comparable to PCA)
    - Canonical Correspondence Analysis (CCA) (comparable to CA)

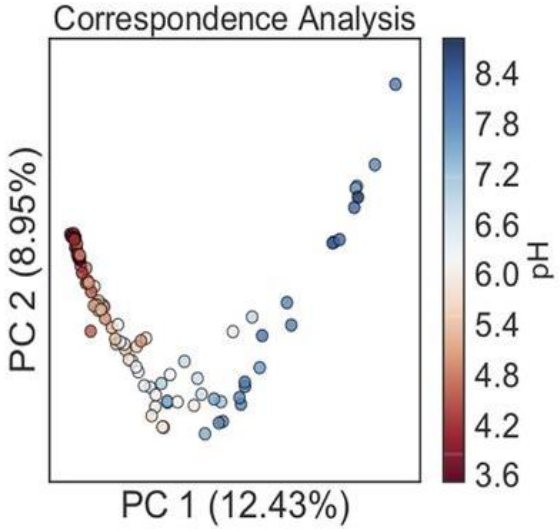
horseshoe/arch effect

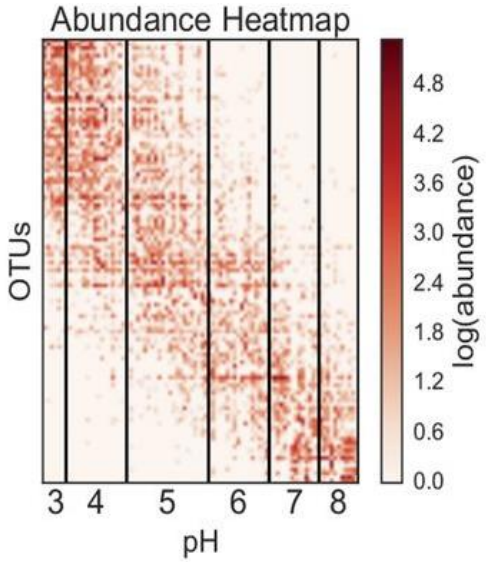
- Artefact of ordination techniques (mainly PCA and CA affected, rare in NMDS)
- Caused by distribution of species along one single gradient



Uncovering the Horseshoe Effect in Microbial Analyses

James T. Morton,[a,b] Liam Toran,[c] Anna Edlund,[d] Jessica L. Metcalf,[e] Christian Lauber,[f] Rob Knight[a,b]



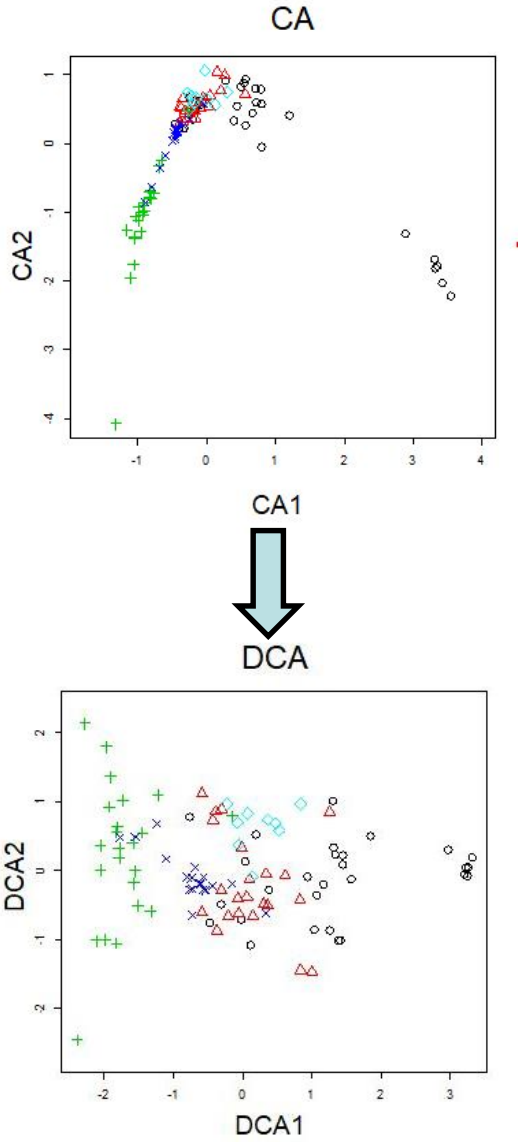


EPFL

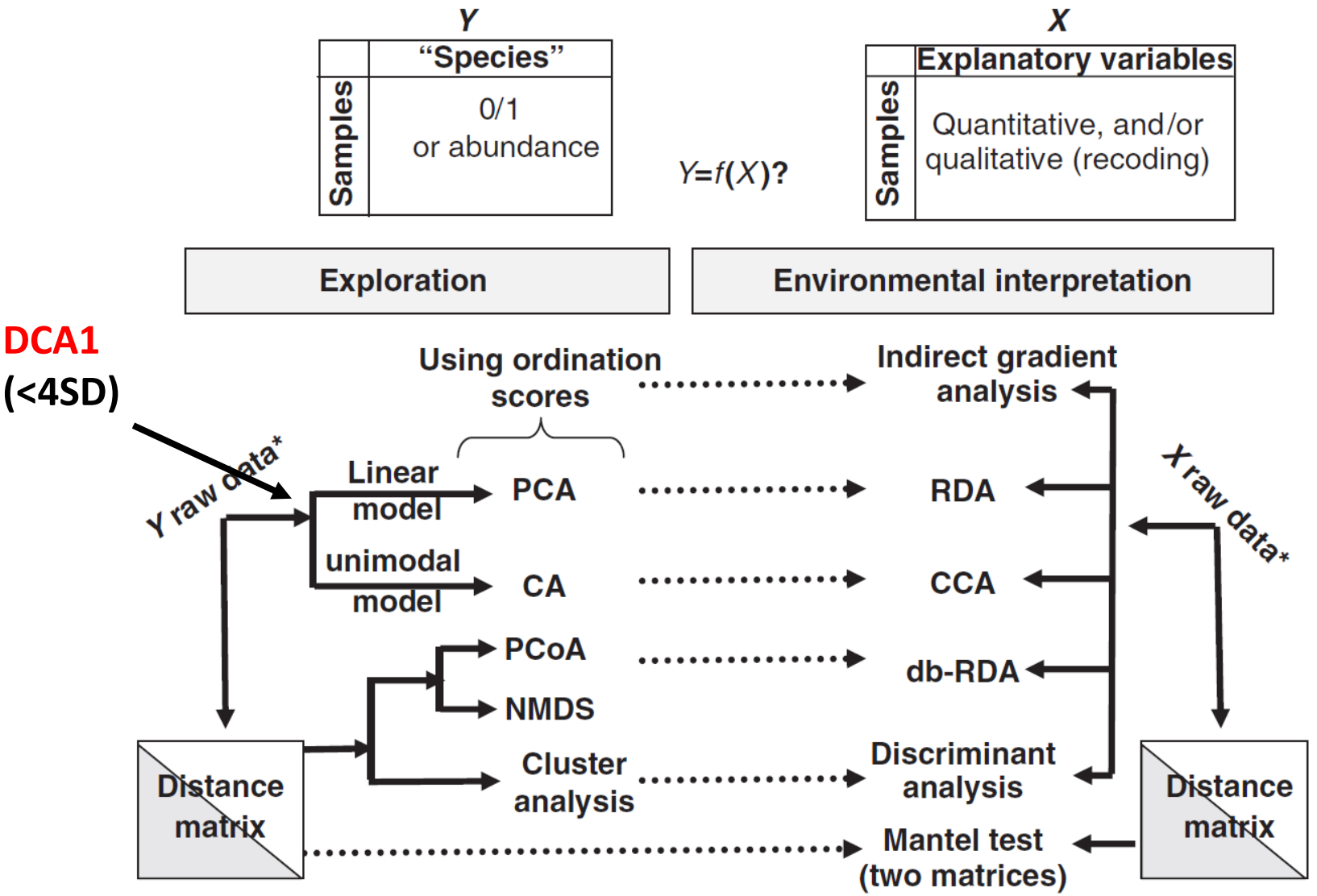
Detrended Correspondance Analysis (DCA)

- removes arch effects by cutting the first axis into segments and shifting sample points along the second axis
- popular method because it often returns meaningful sample distributions
- the length of the first DCA axis (SD of turnover) refers to the heterogeneity (or homogeneity) of the dataset (short vs long ecological gradients)
  - can be used to decide whether data should be analysed by linear (axis shorter than 3 SD, PCA) or unimodal (axis longer than 4 SD, CA) ordination methods

- DCA is criticized and not recommended for use by some of researchers (e.g. Legendre & Legendre 1998, Borcard et al. 2011, or Jari Oksanen (vegan)
  => NMDS is typically more robust!



overview ordinations



Constrained/Canonical Analyses

- Simultaneous analysis of **two datasets**. E.g. species composition (response) and environmental descriptors (explanatory)
- Useful to extract structure of the data that can be interpreted by another dataset
- **Indirect comparison** (Indirect gradient analysis)
  - Correlation of ordination scores (site scores) with explanatory variables (e.g. pH, temp, etc…)
  - Interpretation of an unconstrained ordination
- **Direct gradient analysis**
  - Simultaneous ordination of two datasets (response and explanatory)
  - Analysis/ordination under a constraint (LDA, RDA, CCA, CAP)
    - The ordination axes are forced to express a linear combination of the explanatory variables

Constrained vs unconstrained ordination

- Unconstrained ordination tries to display the main variation in data.
- Constrained ordination tries to display only the variation that can be explained with constraining variables.

=> You will observe patterns you ask for ("constrain") – these are not necessarily the most important!

=> Important to understand how much variation is explained by the constraints.

Principle of constrained ordination

- Extension of (multiple) linear regression to multivariate datasets
- What is the proportion of variation in a set of response variables that can be attributed to a set of explanatory variables?

Data to be explained	Explanatory variables	Type of analysis, statistical model
1 variable (univariate response)	1 variable	Simple linear regression
1 variable (univariate response)	*m* variables	Multiple linear regression
p variables (multivariate response)	*m* variables	**Ordination under constraint** RDA Canonical redundancy analysis CCA Canonical correspondence analysis CAP Canonical analysis of principal coordinates

Redundancy Analysis (RDA)

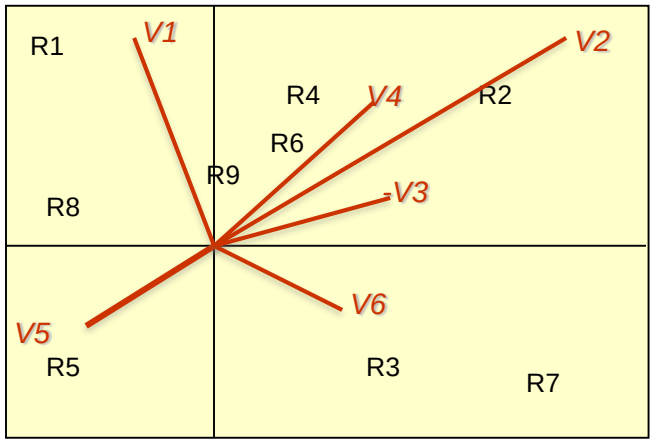
- Extension of PCA
- Applicable when the dataset can be analyzed with PCA:
  - Response variables are in linear relations with each other and with the gradient expressed by the latent variables (components)
  - Euclidean distance is appropriate for measuring the relationships between objects
- The number of explanatory variables must be less than or equal to the number of objects (to avoid overfitting)
- Explanatory variables are automatically standardized and qualitative variables transformed into *dummy variables*

PCA and RDA

Y matrix
(9 objects (R) x
6 variables (V)

PCA

PCA biplot

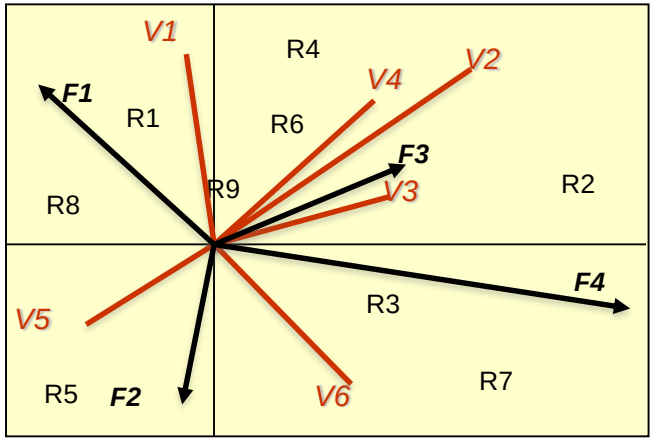


Y matrix
of the response
variables
(9 objects x
6 variables)

X matrix
of the explanatory
variables
(9 objects x
4 variables (F)

RDA

RDA triplot



EPFL

RDA Interpretation

- same principals as for PCA (scaling!)

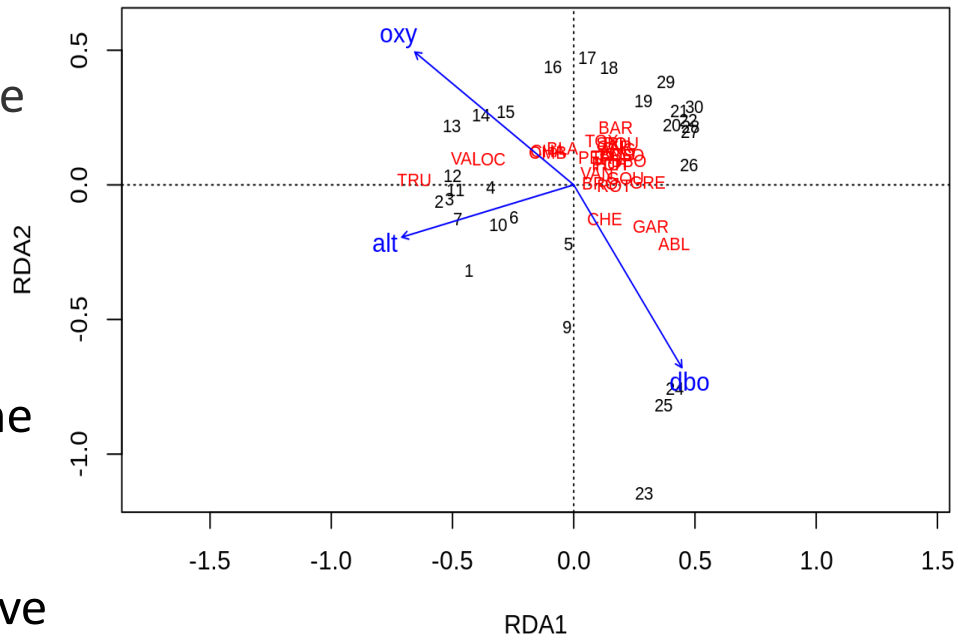
Sites (here numbers) that are close have *similar* communities.

Species (here abbreviations) that are close occupy similar sites.

Arrows show explanatory variables**:**
Longer arrows indicate that the variable *strongly* drives the variation in the community matrix.

Arrows pointing in opposite directions have a *negative* relationship/arrows pointing in the same direction are positively correlated.

RDA triplot



Canonical Correspondence Analysis (CCA)

- Equivalent to CA, but integrates regression of environmental variables
- Same principle as for RDA

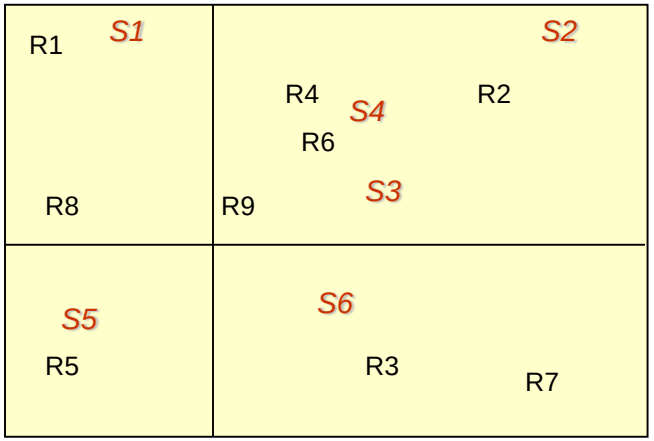
Underlying model (assumption)	Without constraint Unsupervised ordination	With constraint Supervised ordination
Linear (DCA axis <3)	PCA	RDA
Unimodal (DCA axis >4)	CA	CCA

CA and CCA

Matrix **Y**
9 sites (R) x
6 species (S)

CA

CA biplot

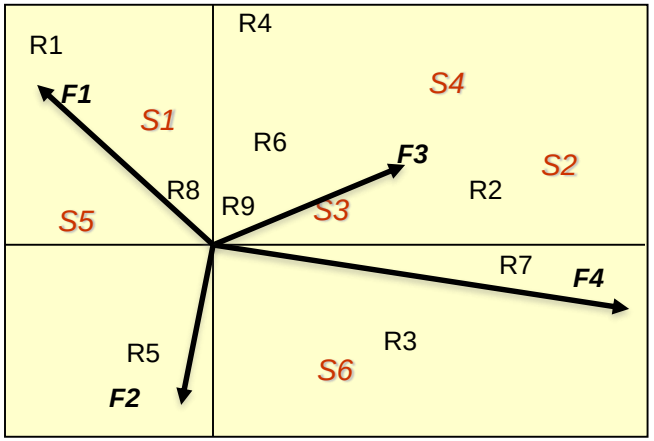


Matrix **Y**
response variables
(9 sites x
6 species)

Matrix **X**
explanatory
variables
(9 sites x
4 variables (F))

CCA

CCA triplot



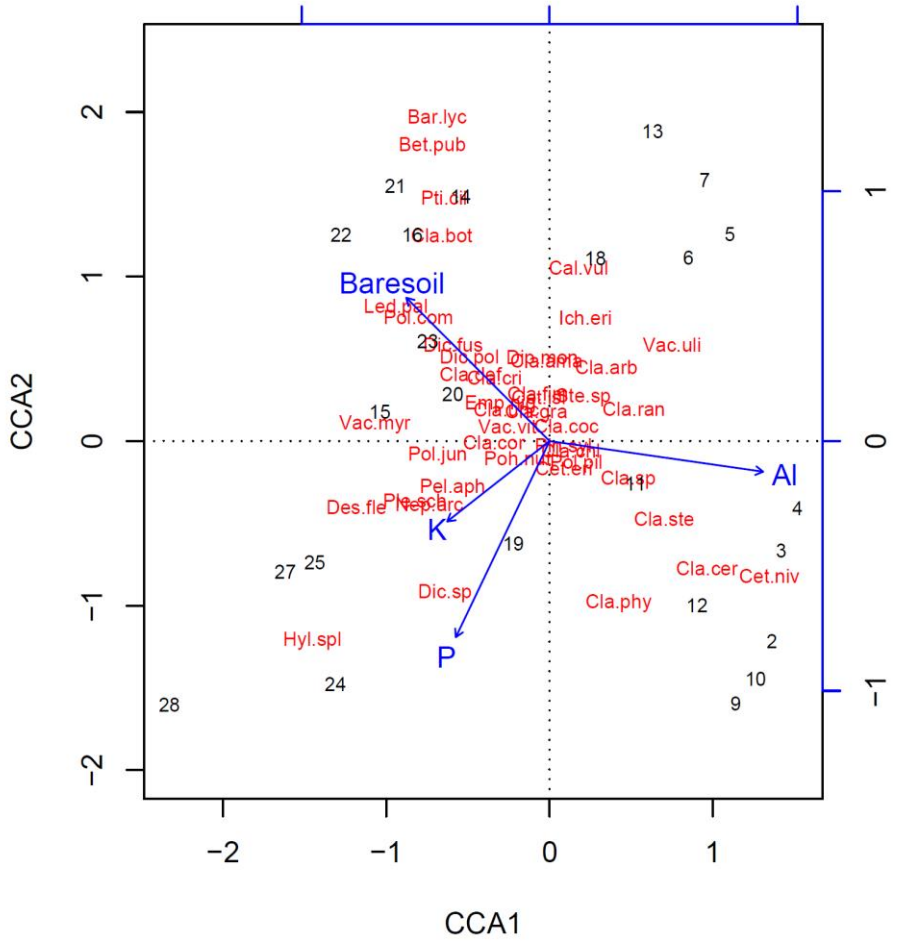
EPFL

CCA Interpretation

- Same principles as for CA
- Additional interpretation of the explanatory variables

Arrows show constraints

Popular to scale by species (scaling 2).



EPFL

Selection of explanatory variables

- **Deductive approach: test of *a priori* hypothesis**
  - The variables are chosen according to hypotheses
  - Possible to account for interactions between explanatory variables

- **Inductive approach (exploratory): no *a priori* hypothesis**
  - Step-by-step selection of explanatory variables (*stepwise selection*)
    - ***Forward selection*** : start from the **null model** (no explanatory variable) and add variables one by one
    - ***Backward selection*** : start from the **full model** (with all explanatory variables) and remove variables one by one
  - Use of optimization criteria (AIC, BIC)

- Examine **variance inflation factor** (VIF) to identify co-linearity between explanatory variables
- Use **adjusted R^2** and **goodness-of-fit** statistics to select explanatory (and response) variables.

  => parsimonious model

Some general advice regarding constrained ordination

- Ordination under constraint of **a single explanatory variable** allows isolating this variable (hypothesis testing)
- The number of explanatory variables should be limited (*forward/backward selection*). Examine the variance inflation factor VIF
- many constraints = no constraints…
- Spatial coordinates or time can be used as explanatory variables or as covariables (can be removed/factored out)