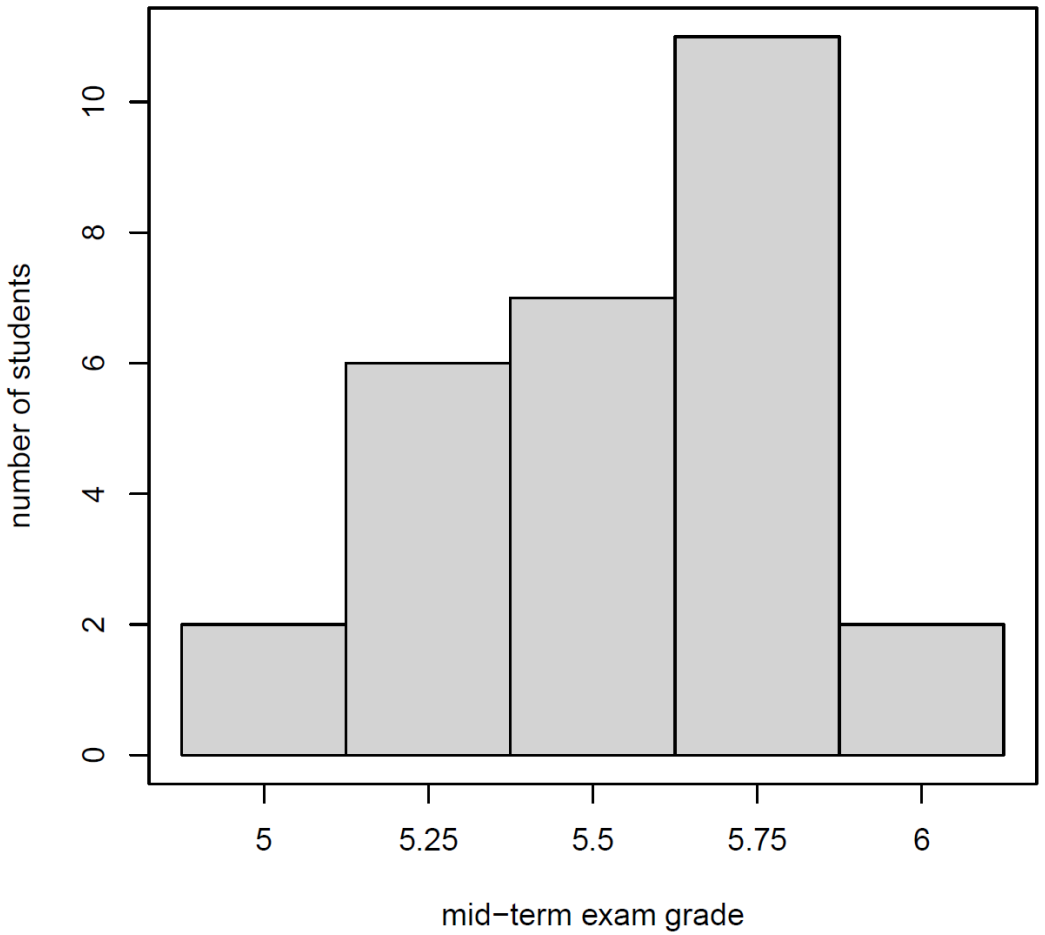
Multivariate statistics in R

Hannes PETER
Martin BOUTROUX

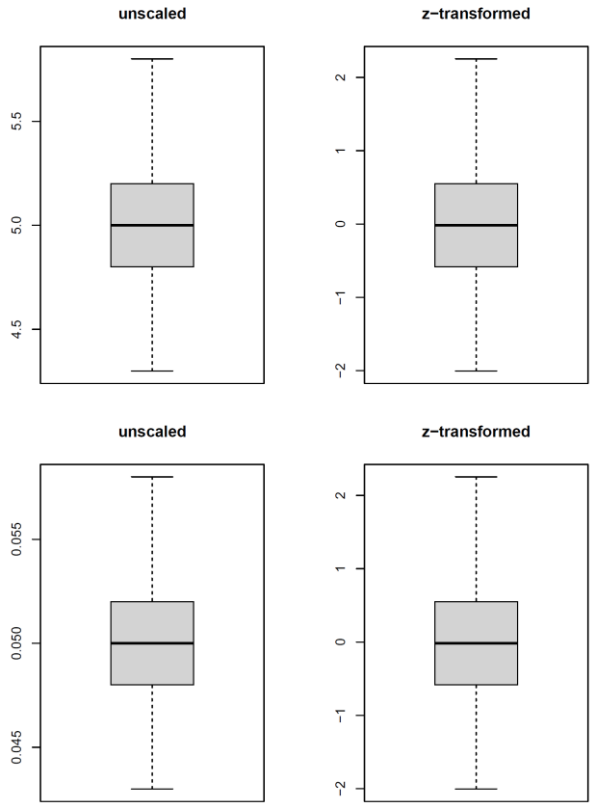
mid-term results



mid-term results

Range does not systematically change!

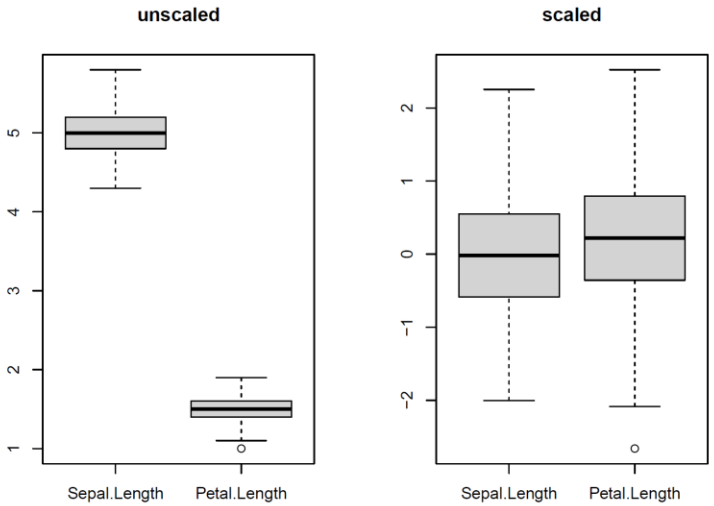
Iris setosa sepal length



Standardization (z-transformation) allows to…:

- ~~compare different variables~~
- improve left-skewed variable distributions
- reduce the range of values
- emphasize the importance of rare taxa

Allows to compare different variables

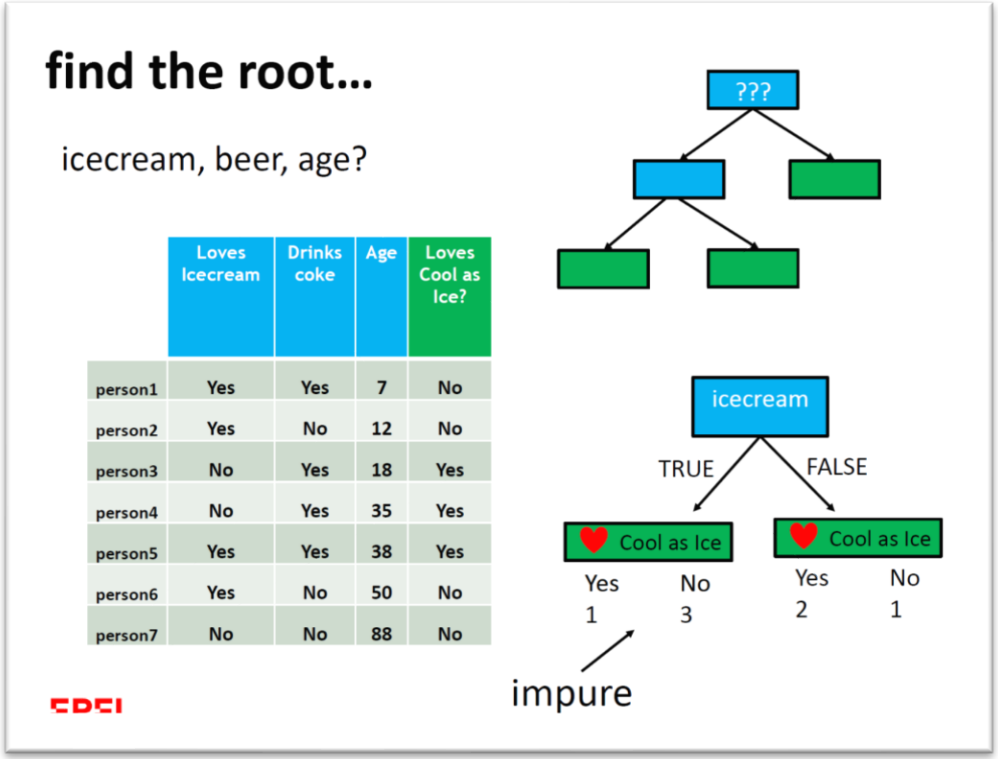


EPFL

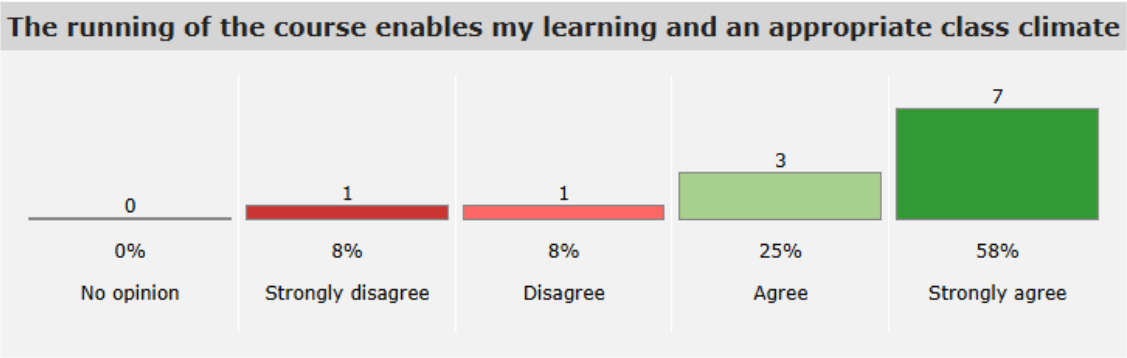
mid-term results

Supervised classification techniques…

- require manual supervision of cluster generation
- build on the fixed assignment of species to objects/samples
- ~~use a multivariate dataset to model additional information (a univariate response)~~
- require an *a priori* defined decision tree (internal structure of dis/similarities)



Indicative feedback



comments:

- systematically lost (R code)
- unclear what the key concepts are
- not enough detail
- lack of clear explanation of statistics
- refresher and more explanation needed
- slides unclear
- no breaks
- the TA is enthusiastic and kind

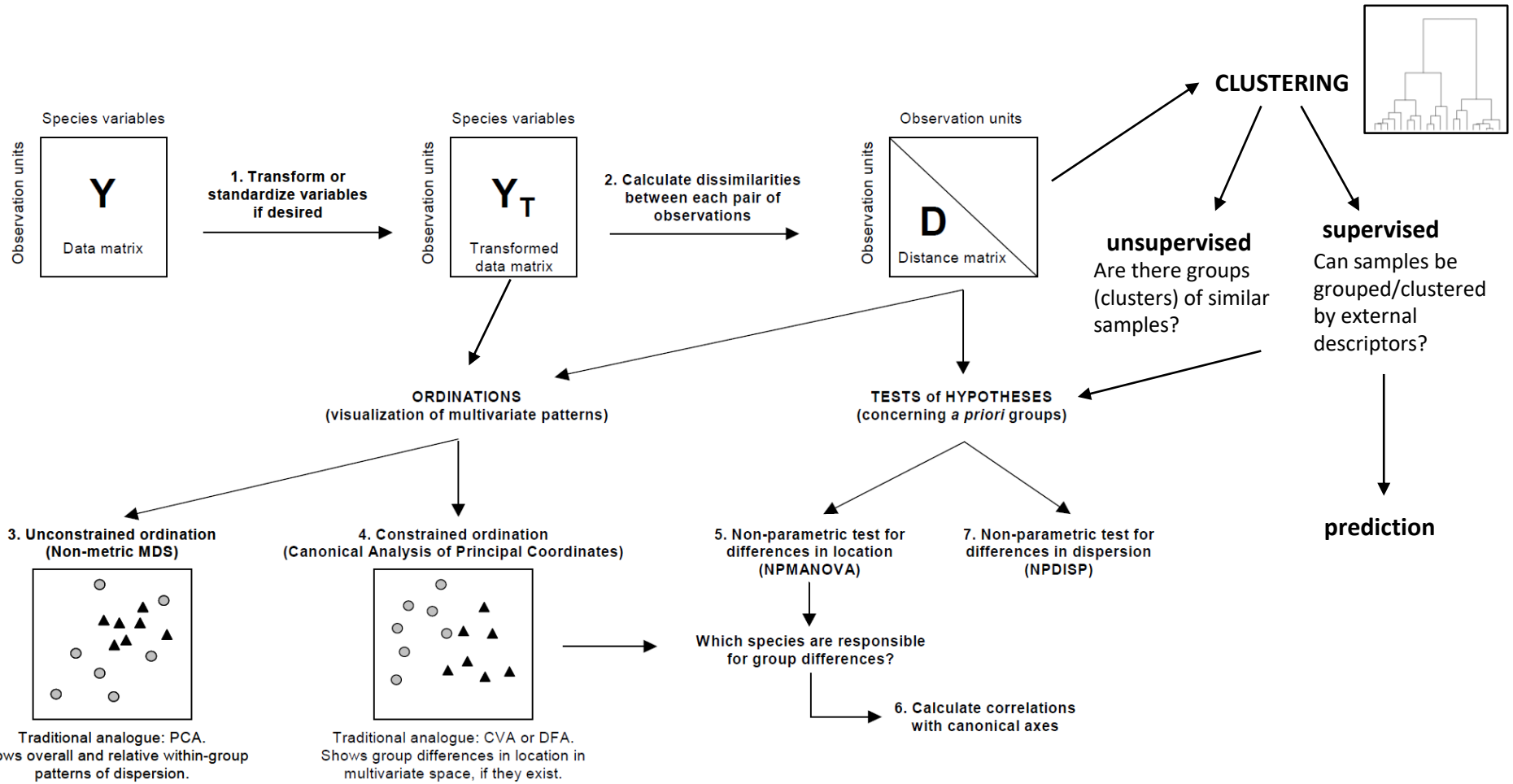
ideas:

- encourage to ask/interrupt if something is unclear
- provide summary & lookout to highlight what is important; interactive
- change title of course:
  - «Applied multivariate statistics using R»
  - «Introduction to Multivariate statistics using R»
  - «Applied Multivariate Statistics for Environmental Scientists»
- more smaller breaks?
- flipped classroom for code part?
- keep Martin

Updated schedule

- 11.09. **session 1** Introduction to multivariate statistics using R
- 18.09. **session 2** Similarity and distance
- 25.09. **group work - define research topic**
- 02.10. **session 3** Cluster analysis
- 09.10. ***«modern R» with Martin*** (tidyverse)
- 16.10. **group work - cluster analysis**
- 23.10. *break*
- 30.10. **session 4** supervised classification
- 06.11. **session 5** unconstrained ordination
- 13.11. **mid-term exam**, **group work - unconstrained ordination**
- 20.11. **session 6** Constrained ordination
- 27.11. Variance partitioning, Auxilliary tools
- 04.12. Diversity, **group work**
- 11.12. **group work**
- 18.12. hand in report
- 08.01. group presentations 1
- 15.01. group presentations 2

recap - ordination

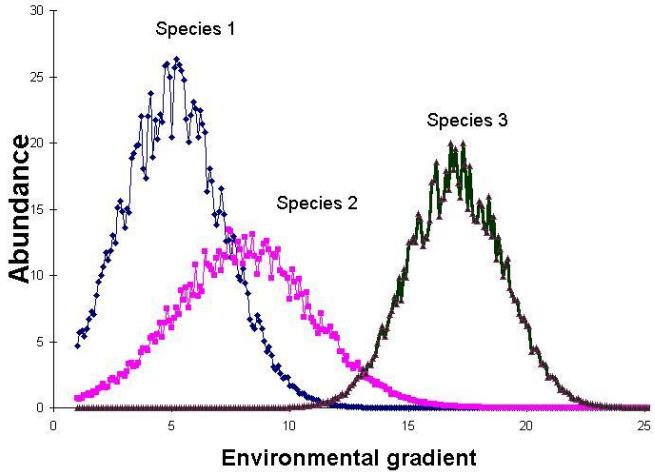


Outlook

- The horseshoe/arch effect
  - An artefact that arises in unconstrained ordination when there is complete turnover in an assemblage along a single environmental gradient
  - Detrended Correspondence Analysis 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
  - Examples of constrained ordination techniques
    - Redundancy Analysis (RDA) (comparable to PCA)
    - Canonical Correspondence Analysis (CCA) (comparable to CA)

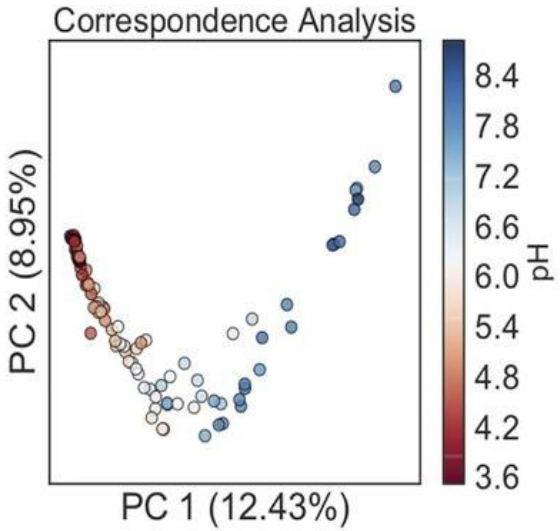
horseshoe/arch effect

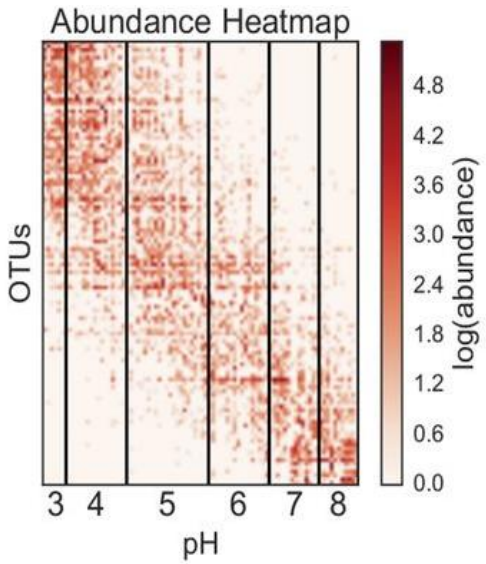
- Artefact of ordination techniques (mainly PCA and CA affected, rare in NMDS)
- Caused by distribution of species along one single gradient



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]



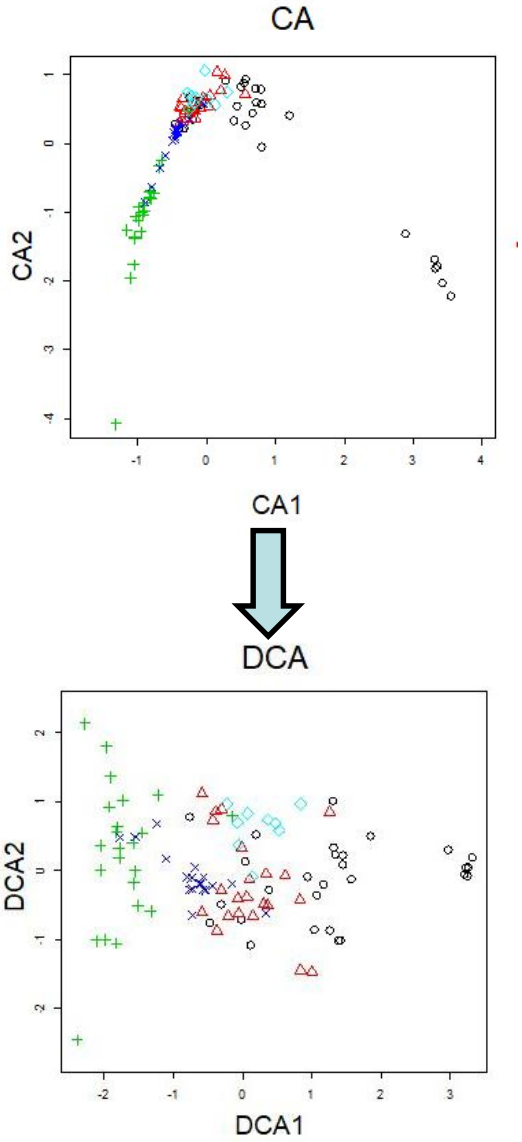


EPFL

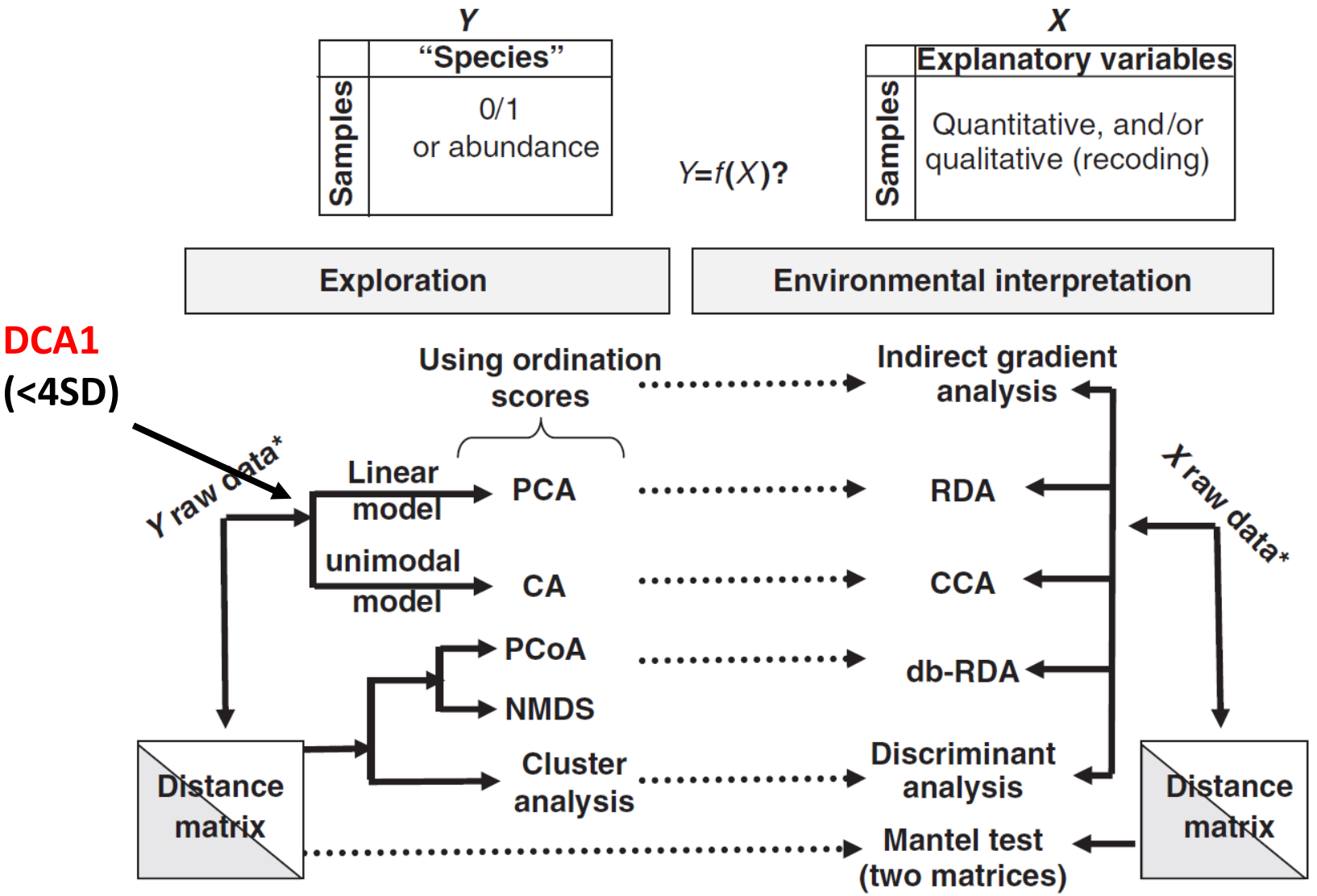
Detrended Correspondance Analysis (DCA)

- 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 gradients)
  - can be used to decide whether data should be analysed by linear (axis shorter than 3 SD) or unimodal (axis longer than 4 SD) 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 can observe only things you have measured...

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)

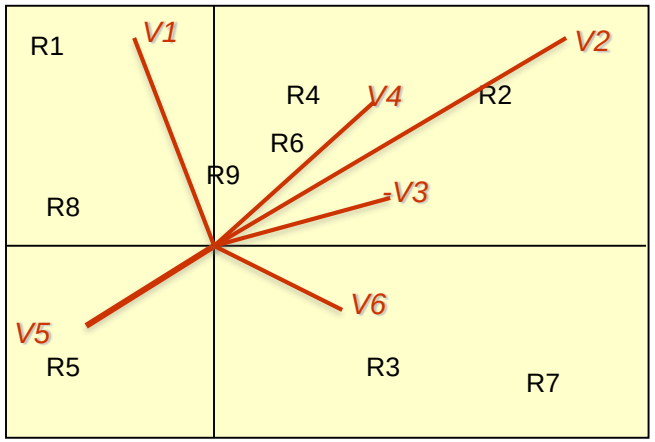
- Extension of PCA
- RDA axes are constrained such that the first axes (canonical axes) are linear combinations of the explanatory variables
- Applicable when the dataset can be analyzed with PCA:
  - Response variables are in linear relation with each other and with the gradient expressed by the latent 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

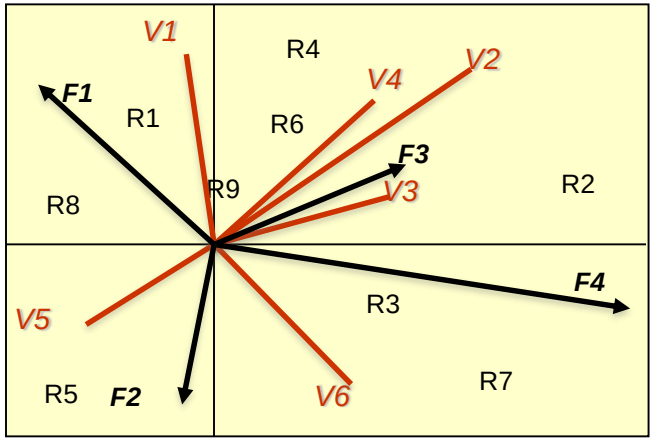


Y matrix
of the response
variables
(9 objects x
6 variables)

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

RDA

RDA triplot



EPFL

RDA Interpretation

- same principals as for PCA (also choice of scaling)

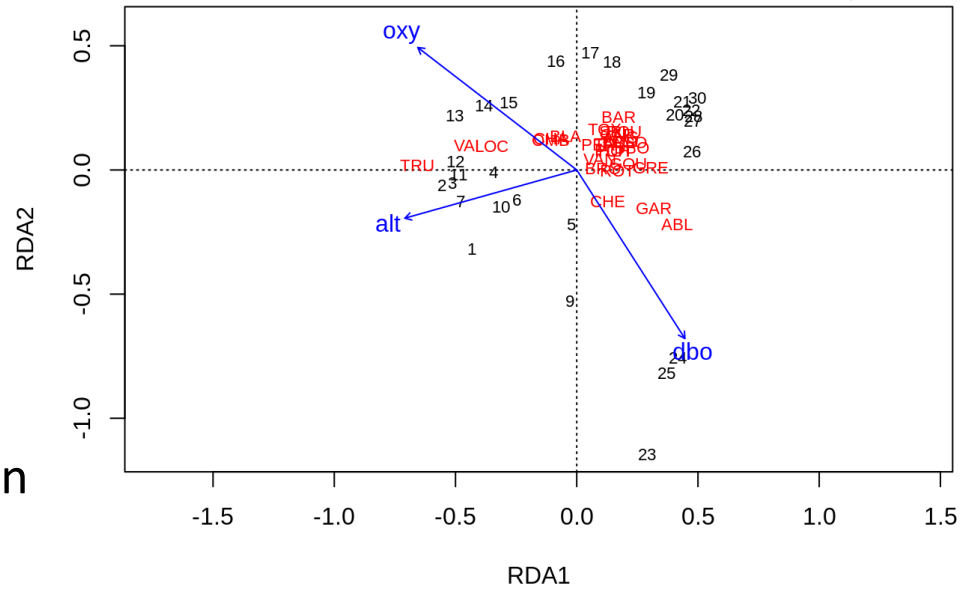
Sites (here numbers) that are close have *similar* communities.

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



EPFL

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
Unimodal	CA	CCA
Linear	PCA	RDA

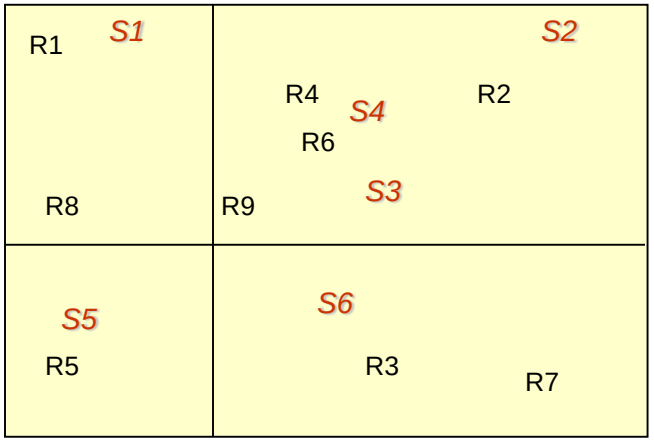
(see DCA axis)

CA and CCA

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

CA

CA biplot

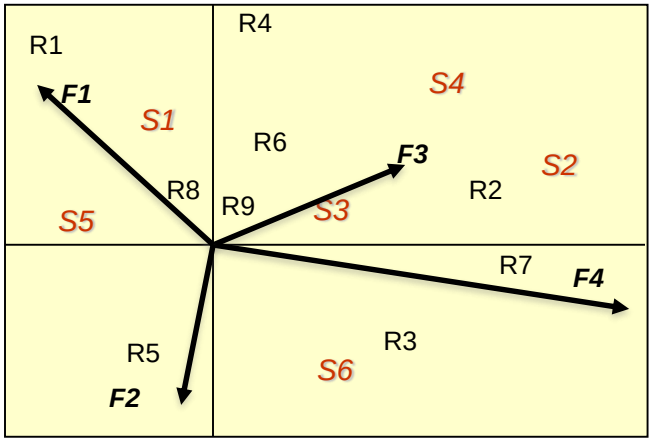


Matrix **Y**
response variables
(9 sites x
6 species)

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

CCA

CCA triplot



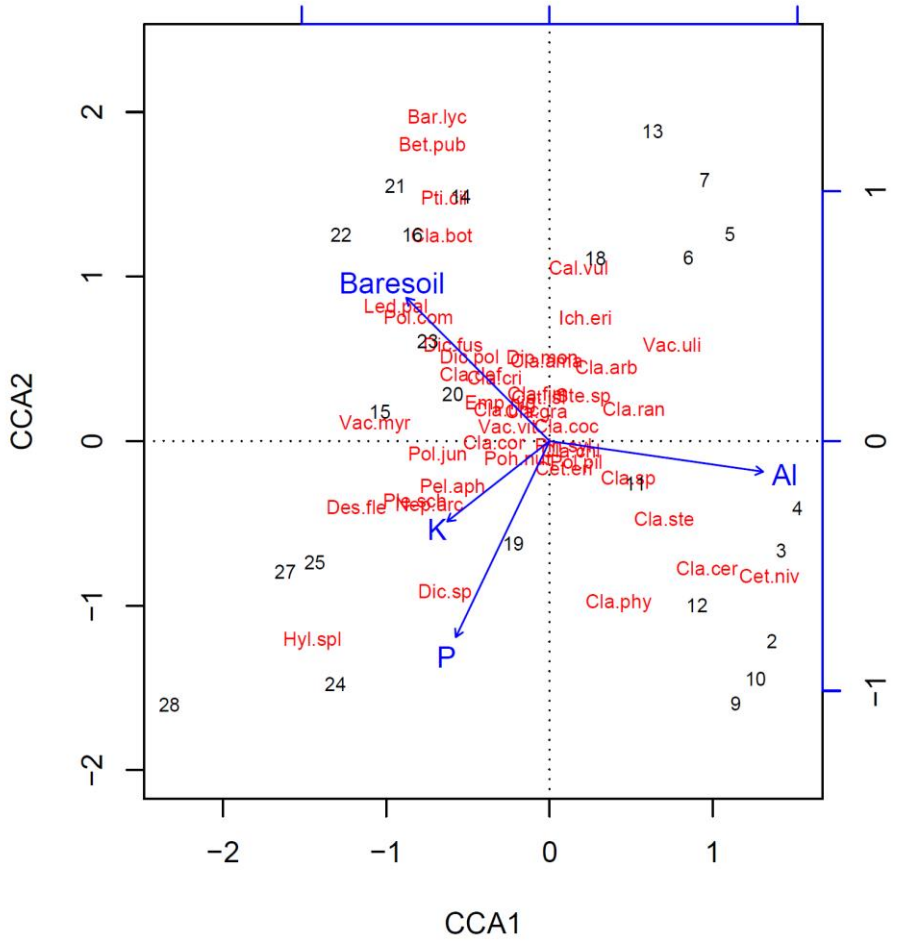
EPFL

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
  - Possibile 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)

Take home messages