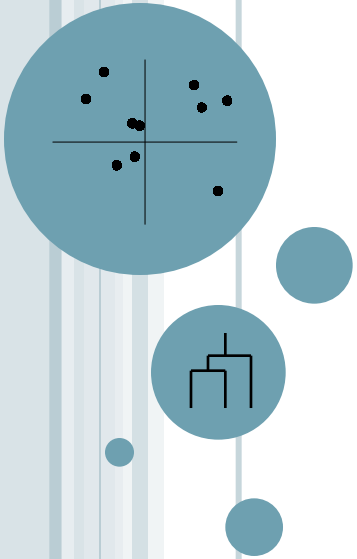


Introduction to multivariate analysis of ecological data

David Zelený & Ching-Feng Li

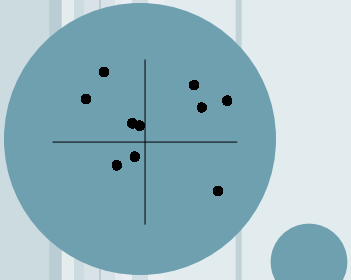
VEGETATION SCIENCE GROUP
MASARYK UNIVERSITY BRNO

Introduction to multivariate analysis

- Ecologial similarity
  - similarity and distance indices
- Gradient analysis
  - regression, calibration, ordination
  - linear and unimodal methods
  - unconstrained and constrained ordination
  - eigenvalue-based and distance-based ordinations
  - partial ordination, variation partitioning
  - Monte-Carlo permutation tests, forward selection
- Classification
  - hierarchical and non-hierarchical algorithms
  - cluster analysis, TWINSPAN

LITERATURE

- Gotelli & Ellison (2004) **A Primer of Ecological Statistics.** Sinauer Associates.
  - well written, excellent for beginners; not too much about multivariate analysis
- Lepš & Šmilauer (2003) **Multivariate Analysis of Ecological Data Using CANOCO**. Cambridge.
  - less theory, more practical use, focused on CANOCO users, case studies for independent work including training datasets
- Zuur, Ieno & Smith (2007) **Analysing Ecological Data.** Statistics for Biology and Health. Springer.
  - well explained basics of various methods used for analysis of ecological data, clever examples
- Legendre & Legendre (1998) **Numerical Ecology.** 2nd English Edition, Elsevier.
  - bible for numerical ecology, surprisingly also quite readable
- Ordination website of Mike Palmer (http://ordination.okstate.edu/)
  - comprehensive introduction, but a bit out of date



ECOLOGICAL SIMILARITY

Similarity and distance indices

Similarities × Distances

Similarity indices

- represent the similarity between samples, not their position in multidimensional space
- lowest value 0 – samples have no species in common
- highest value (1 or other) – samples are identical

Distances among samples

- allows to locate the sample in multidimensional space
- the lowest value 0 – samples are identical (at the same location)
- value increases with the increasing dissimilarity between samples

Problem of „double-zeros”

The fact, that species is missing simultaneously in both samples (double-zeros), can have several meanings:

- on the gradient, samples are located outside the species ecological niche
  - but we cannot say if both samples are located at the same end of ecological gradient (and are thus quite similar), or they are located on opposite sites of the gradient (and thus they are quite different)
- on the gradient, samples are located inside the species ecological niche, but the species is missing, because
  - it didn’t get there (*dispersal limitation*)
  - we overlooked it (*sampling bias*)
  - just now it’s in dormant stage and we cannot see it (therophytes, geophytes)

Problem of „double-zeros”

	wet species 1	wet species 2	mesic species 1	mesic species 2	dry species 1	dry species 2
sample 1	1	1	0	0	0	0
sample 2	0	1	1	1	1	0
sample 3	0	0	0	0	1	1

- **symetrical indices of similarity:** double zeros in two samples increase similarity of these samples
- **asymetrical indices:** double zeros are ignored

Similarity indices

qualitative × quantitative

- qualitative – for presence-absence data
- quantitative – for counts, abundances etc.

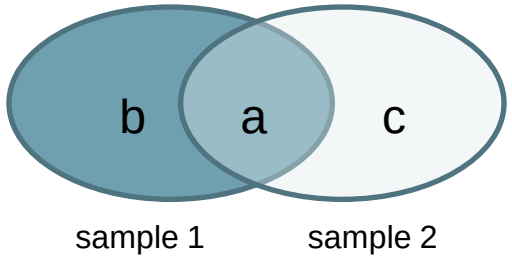
symmetrical × asymmetrical

- symmetrical – treats double-zeros in a same way as double-presences (they contain information about similarity of samples); rarely used in community ecology
- asymmetrical – ignores double-zeros; the most common indices in community ecology

metrics × semimetrics

- semimetrics do not follow triangle inequality rule and cannot be used to order points in Euclidean (metric) space

SIMILARITY INDICES



Qualitative data (presence-absence)

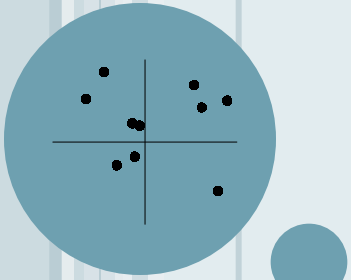
- Jaccard index $J = a / (a + b + c)$
- Sørensen index $S = 2a / (2a + b + c)$
  - presence of the species in both samples (*a*) has double weight compared to Jaccard index
- Simpson index $Si = a / [a + \min (b,c)]$
  - suitable for samples with very different number of species

Quantitative data (cover, abundance)

- e.g. generalized Sørensen index (*percentage similarity*)
  - quantitative variant of Sørensen index
  - suitable for ecological data
  - *percentage dissimilarity (PD,* ***Bray-Curtis index****)* = 1 – PS

DISTANCE MEASURES

- Euclidean distance
  - range of values is strongly dependent on used units
  - intuitive, but sensitive for outliers – not too suitable for ecological data
- chord distance, relativized Euclidean distance
  - Euclidean distance calculated on samples standardized *by sample norm*
- chi-square distance
  - usually not explicitly calculated
  - distance among samples in unimodal ordination techniques (e.g. correspondence analysis)
- all similarity indices could be transformed into distances
  - $D = 1 - S$, or $D = \sqrt{(1 - S)}$
  - square-root formula used e.g. for Sørensen index

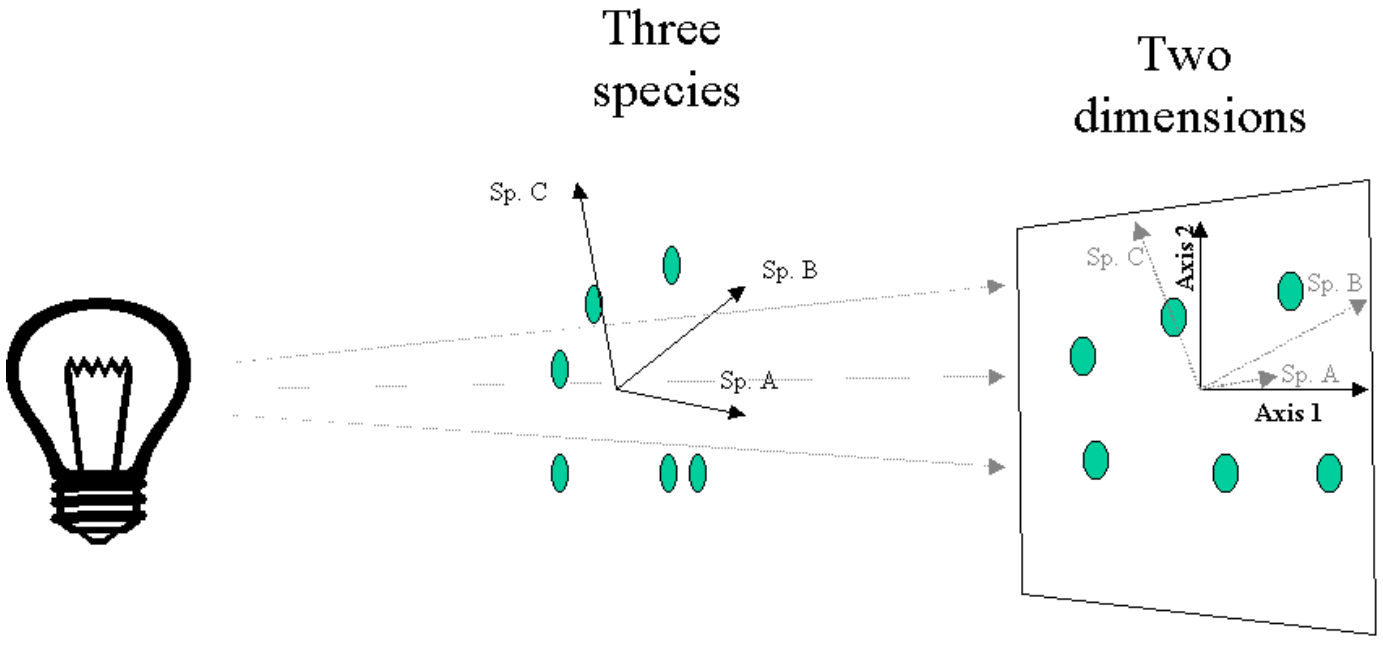


GRADIENT ANALYSIS

ORDINATION

JUSTIFICATION OF THE PROBLEM

- ecological gradient usually influences response (abundance) of several species simultaneously
  - ➔ species data are **redundant** – if I know response of one species, I can (somehow) predict also the behavior of other species
  - ➔ thanks to this redundancy it makes sense to reduce many dimensions of multidimensional space (spaces no. 1-4) into few dimensions of ecological space (space no. 5)
- ➔ if the species response completely independently on each other, ordination (reduction of multidimensional space) is not worth trying – it doesn't bring anything new



http://ordination.okstate.edu/

ORDINATION

FORMULATIONS OF THE PROBLEM

1) **hidden variables (ordination axes) –** find hidden ('latent') variables, that represents the best predictors for the values of all the species
2) **configuration of samples in ordination space –** find such configuration of samples in reduced ordination space, so as their distances in this space correspond to their compositional dissimilarity
3) **reduction of dimensionality –** project multidimensional space defined by particular species into few-dimensional space defined by ordination axes

Unconstrained × constrained ordination

Unconstrained ordination (indirect gradient analysis)

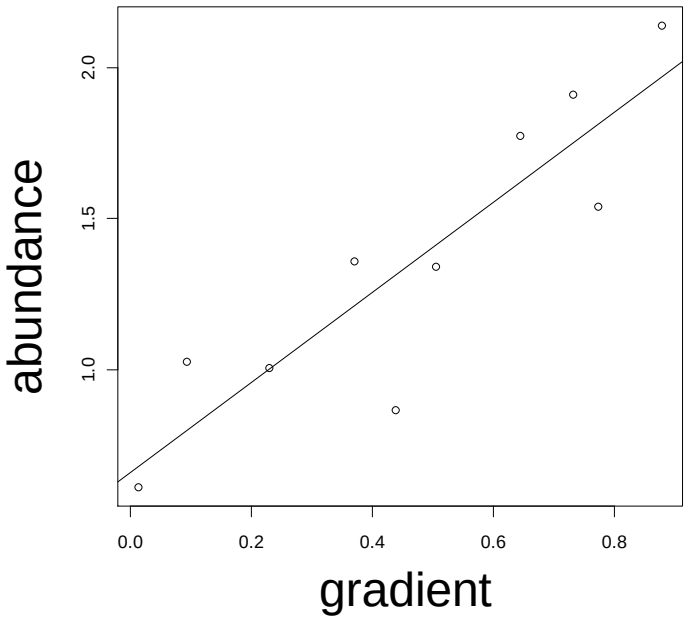
- uses only matrix **samples × species**
- searches for hidden variables (ordination axes), which best represent the variability in species data
- more for hypothesis generation, not testing

Constrained ordination (direct gradient analysis)

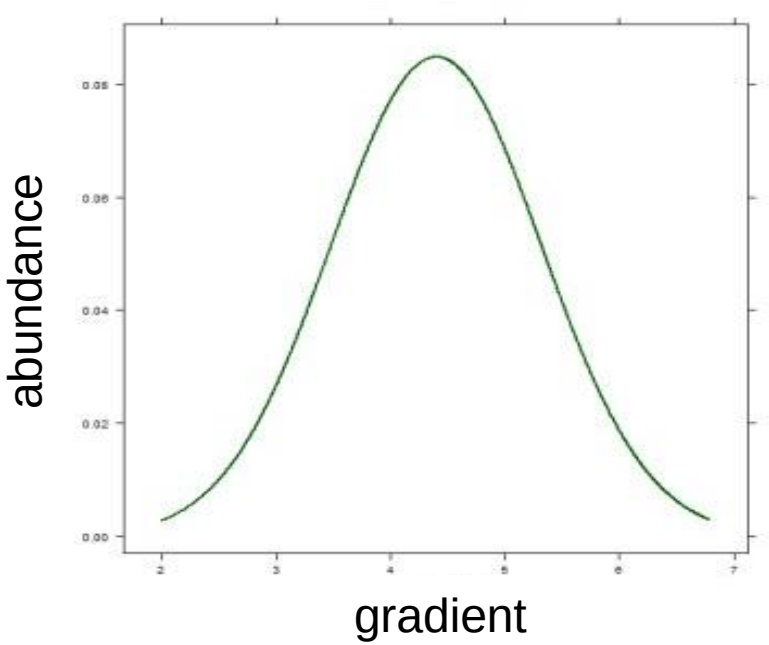
- needs two matrices: **samples × species** and **samples × environmental variables**
- constrained ordination axes represent the directions of the variability in species data, which can be explained by known environmental variables
- more for hypothesis testing than generating

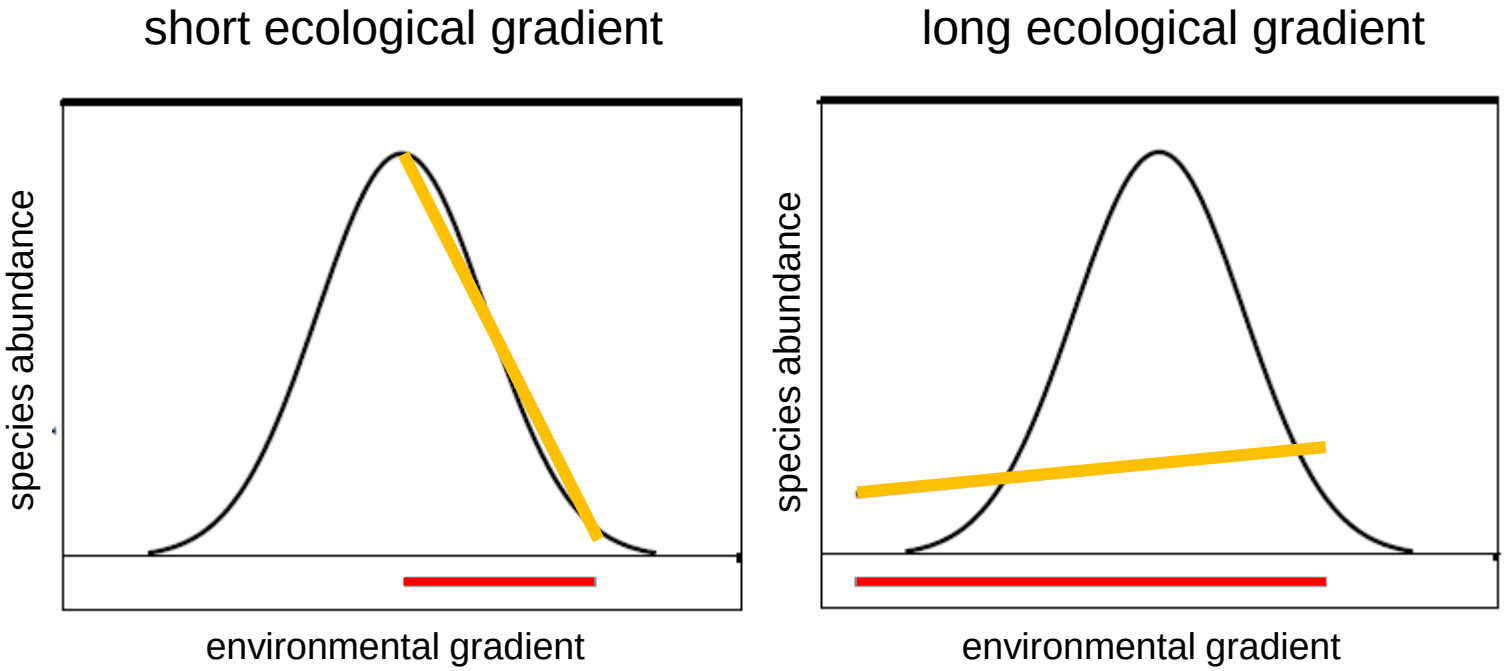
Species response on ecological gradient

linear



unimodal

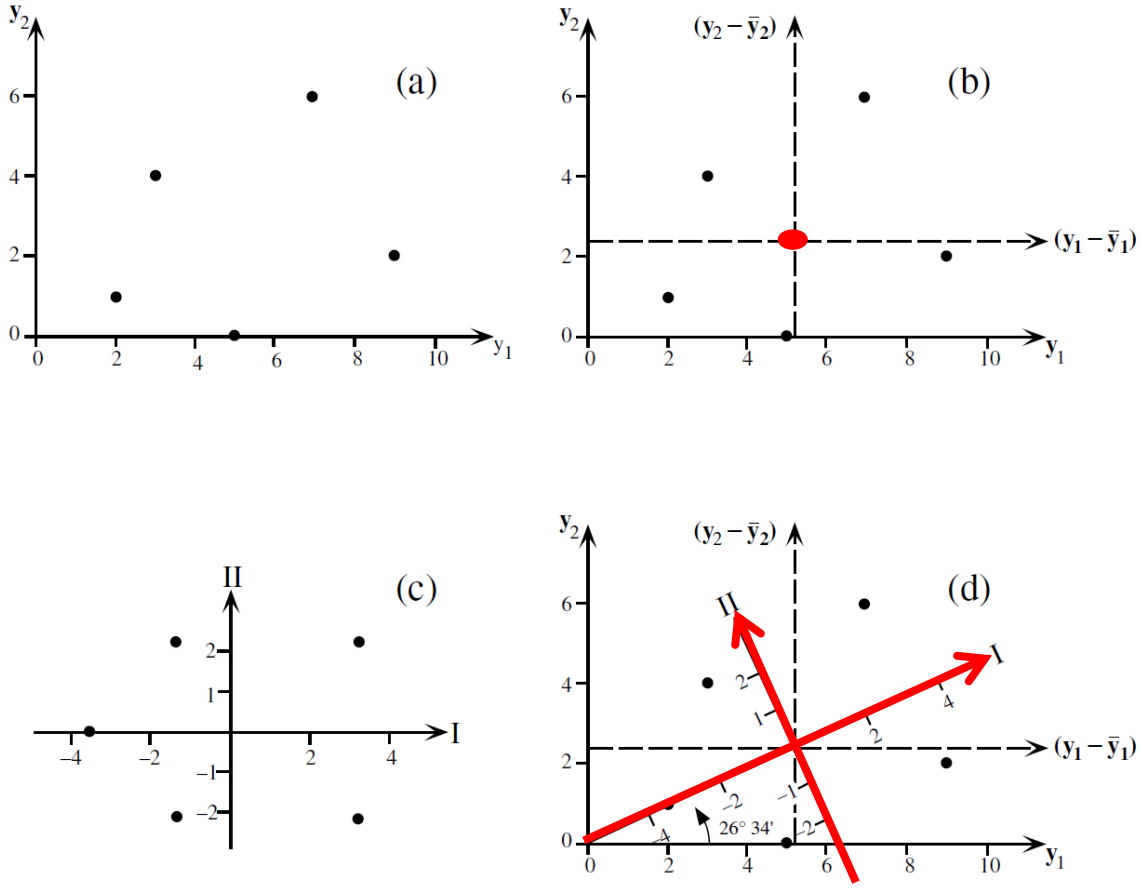



short ecological gradient
species abundance
environmental gradient
long ecological gradient
species abundance
environmental gradient

Basic ordination techniques

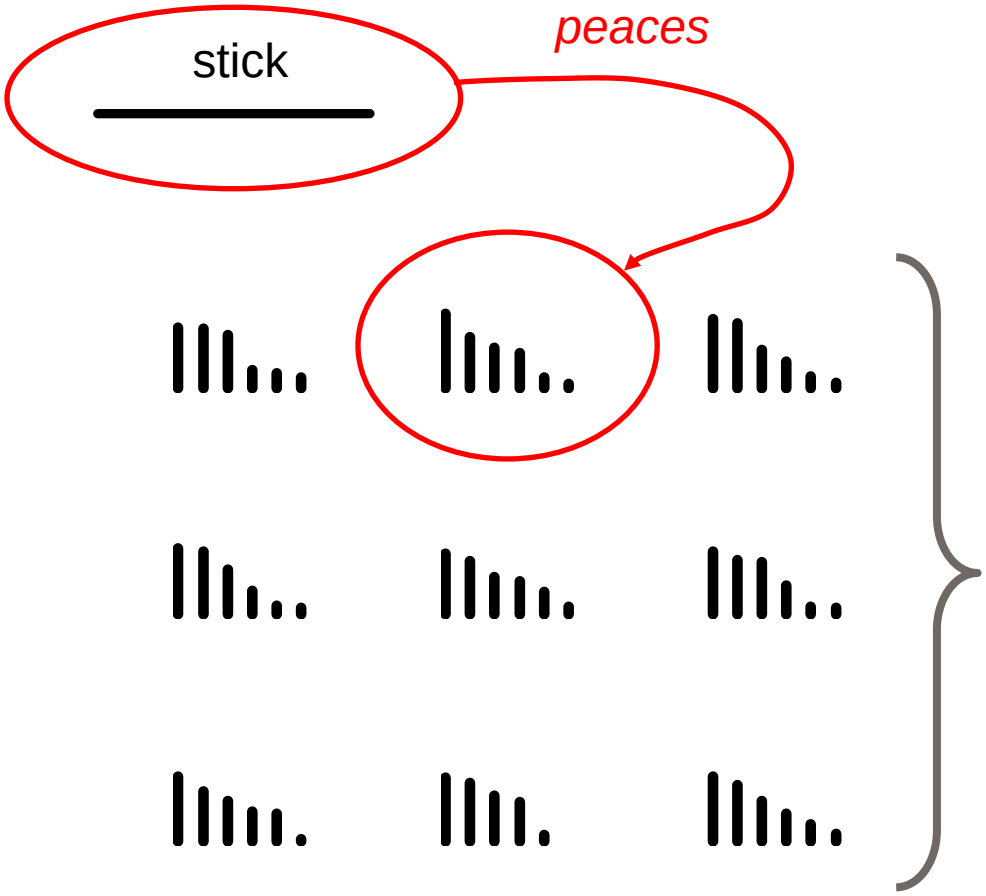
	linear species response	unimodal species response
unconstrained ordination	**PCA** (*Principal Component Analysis*)	**CA** (*Correspondence Analysis*)
constrained ordination	**RDA** (*Redundancy Analysis*)	**CCA** (*Canonical Correspondence Analysis*)

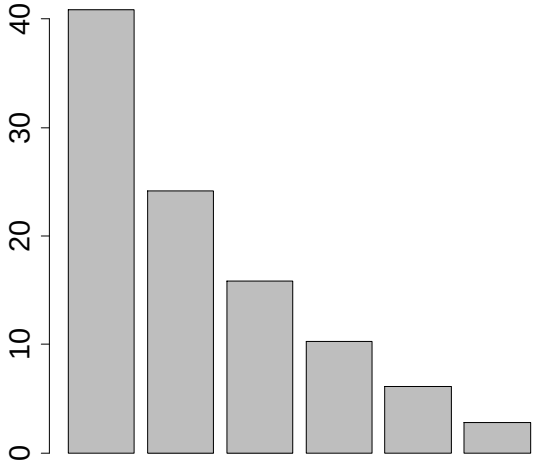
PCA (Principal Component Analysis)



BROKEN-STICK MODEL

stick will break randomly into 6 peaces





EXPLANATORY VARIABLES IN ORDINATION

TWO ALTERNATIVE APPROACHES

1. **unconstrained ordination + correlation**
   - get samples scores on main ordination axes
   - correlate these samples scores with environmental variables

   \+ for sure will catch the main gradients in species composition

   − may not catch the part of variability, which is directly related to measured environmental factors

2. **constrained ordination**
   - environmental variables enter the ordination as explanatory variables
   - sample scores on the ordination axes is directly influenced by these variables

   \+ for sure will catch the part of the variability in species composition, which is directly related to measured environmental factors

   − may lost information about variability in data, which is not directly related to any environmental factor

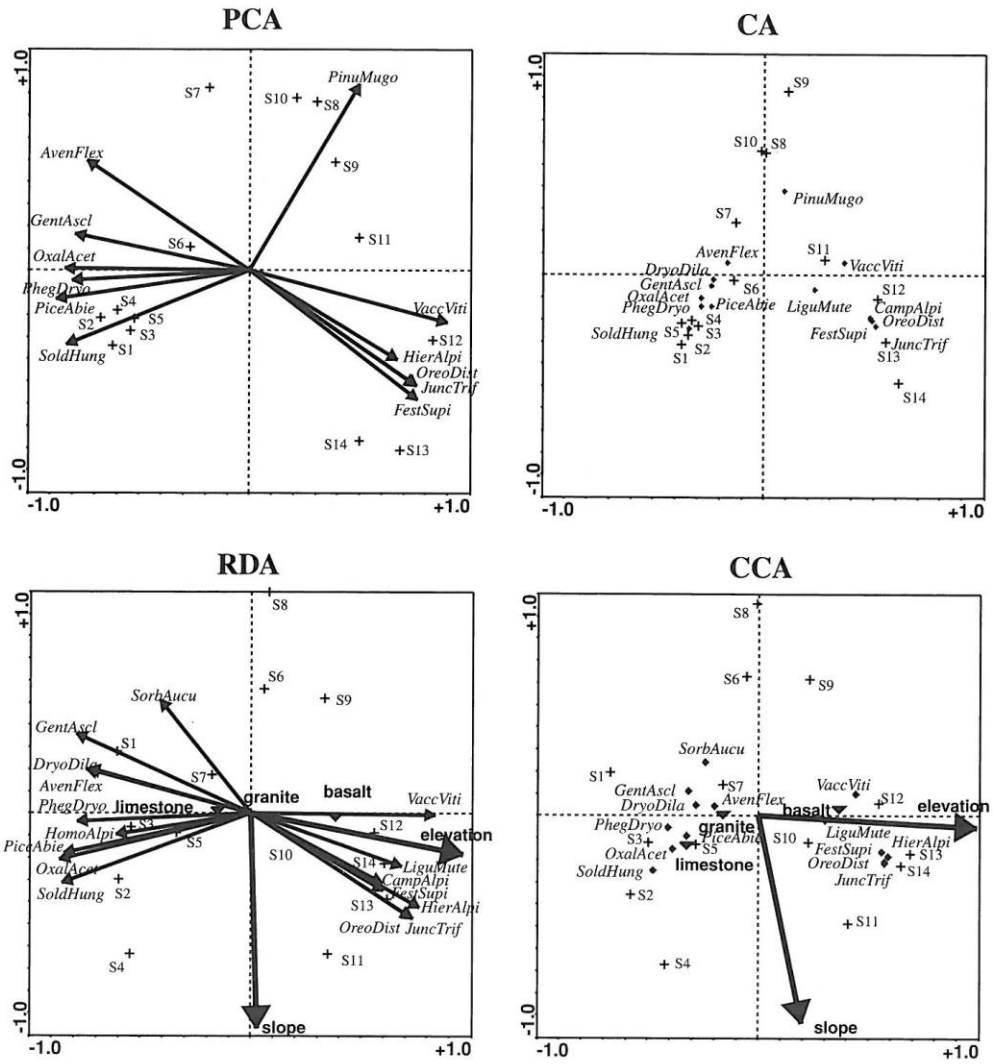
ORDINATION DIAGRAM

linear method

unimodal method

unconstrained ordination

constrained ordination

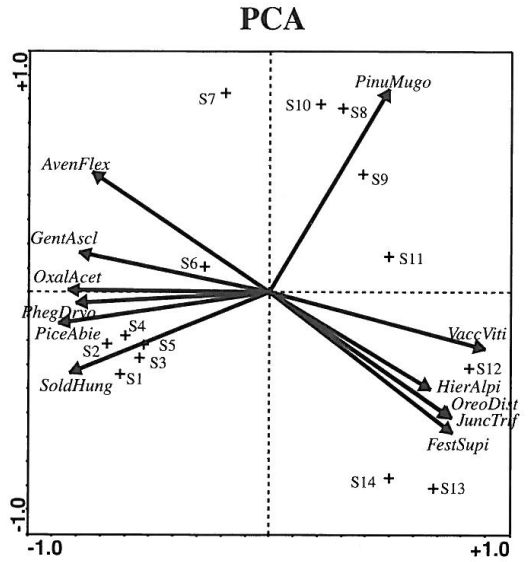


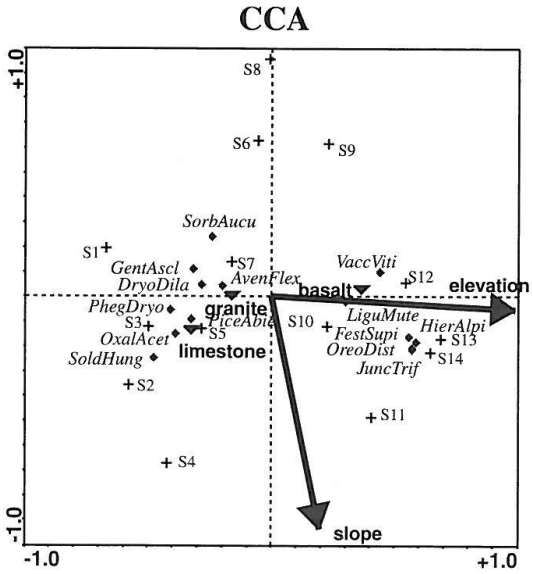
Lepš & Šmilauer (2003) Multivariate analysis of ...

ORDINATION DIAGRAM

RULES FOR VISUALIZATION

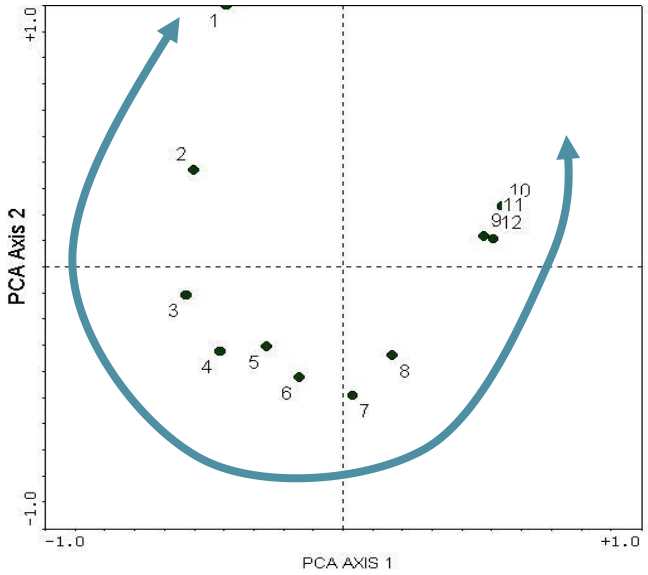
- display of samples
  - -> points
- display of species
  - -> arrows (linear methods)
  - -> points, centroids (unimodal methods)
- display of ordination axes
  - horizontal axis should be axis of higher rank
  - axis orientation is arbitrary
- display of environmental variables
  - arrows (quantitative variables)
  - centroids (categorical variables)
- types of ordination diagrams
  - **scatterplot** - 1 type of data (samples or species)
  - **biplot** - 2 types of data (e.g. samples and species)
  - **triplot** - 3 types of data (samples, species and environmental variables)

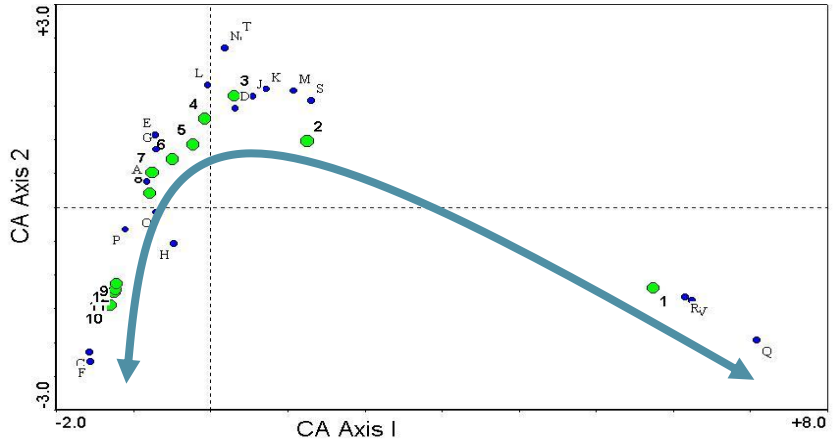




Lepš & Šmilauer (2003) Multivariate analysis of ...

ARTIFACTS IN ORDINATIONS





http://ordination.okstate.edu

Horseshoe effect

- Principal Component Analysis (PCA)
- order of samples along the first axis doesn't reflect their real dissimilarity
- in extreme case, the ends of the horseshoe may cross

Arch effect

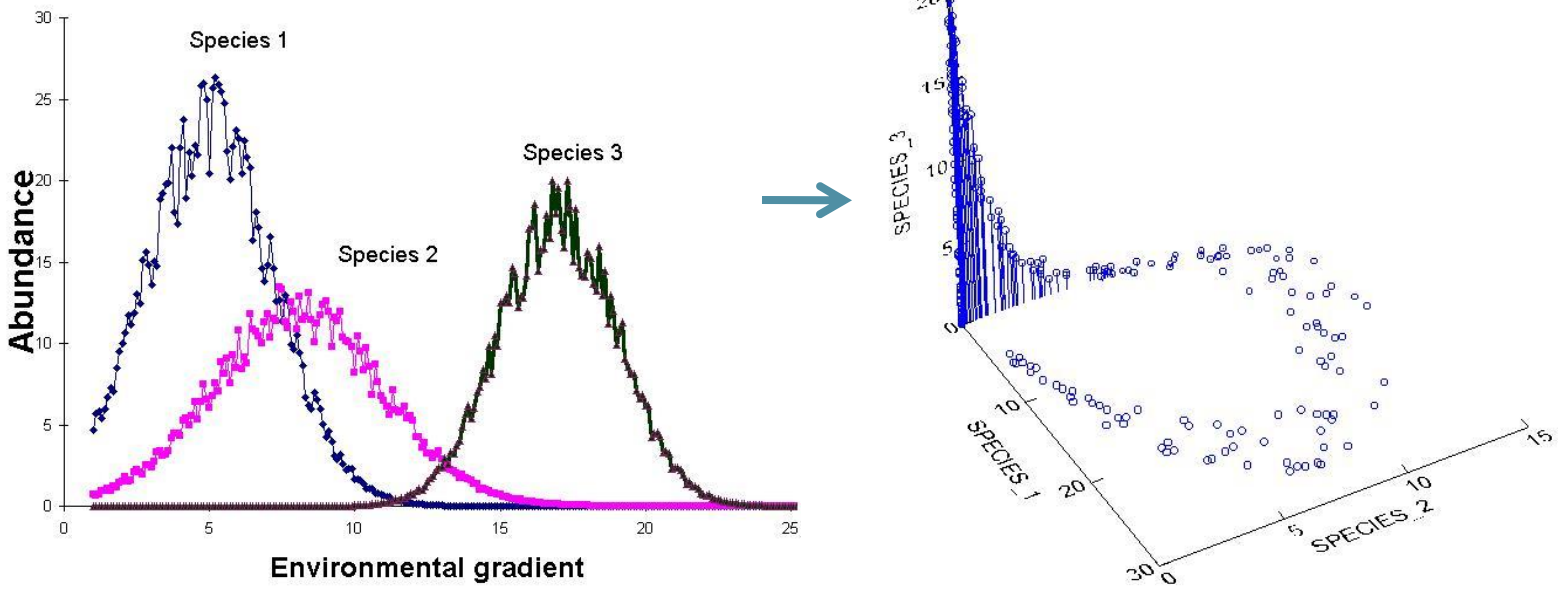
- Correspondence Analysis (CA)
- order of samples along the first axis still reflect their real dissimilarity
- the second axis is non-linear combination of the first axis

Artifacts in ordinations

Horseshoe and arch effects

Possible explanations:

- algorithm consequence – each higher ordination axis has to be linearly independent on the lower one, but nonlinear dependence is not considered
- projection consequence – nonlinear relationships between species and environmental gradients are projected into linear space defined by Euclidean distances



Artifacts in ordinations

Possible solutions

- detrending – removal of the trend from ordination axes
  - *Detrended Correspondence Analysis* (DCA, Hill & Gauch 1980)
  - *detrending by segments* (the most common)
  - *detrending by polynomials* (if there are covariables in analysis)

- use of distance-based ordination techniques, which allows to ordinate the samples using distance coefficients different from Euclidean distance (PCA) or chi-square distance (CA)
  - *Principal Coordinate Analysis* (PCoA, synonym for *Metric Dimensional* Scaling, MDS)
  - *Non-metric Multidimensional Scaling* (NMDS)

Distance-based × eigenvalue-based ordination methods

Distance -based ordination methods

- NMDS and PCoA
- based on dissimilarity matrix
- interpretation is focused on the distance among samples in ordination space

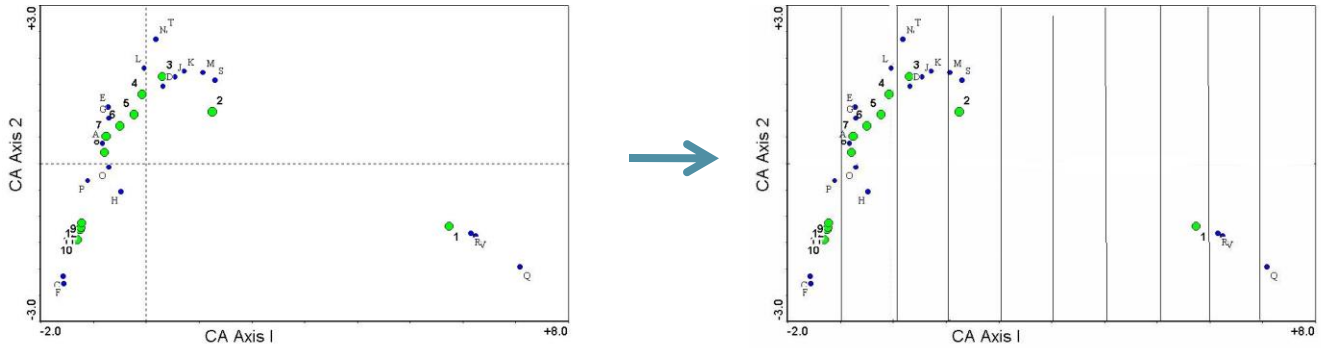
Eigenvalue-based ordination methods

- PCA, CA or DCA, respectively, and their constrained twins RDA, CCA or DCCA, respectively
- based on the matrix samples x species, from which main ordination axes (*eigenvectors*) are extracted
- interpretation is focused on the directions of variability in species data, expressed by particular ordination axes

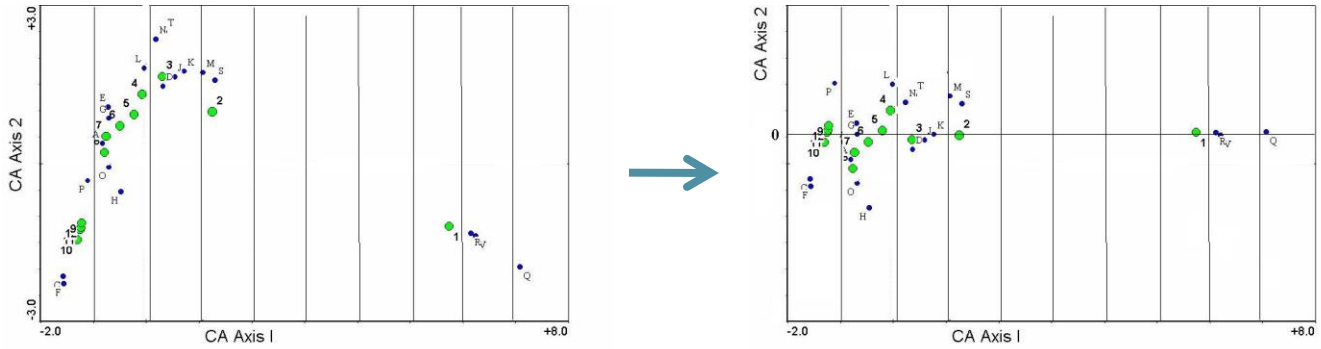
Detrended Correspondence Analysis

The process of "detrending"

Step 1 – the first axis is divided into several segments



Step 2 – samples in each segment are centered along the second axis



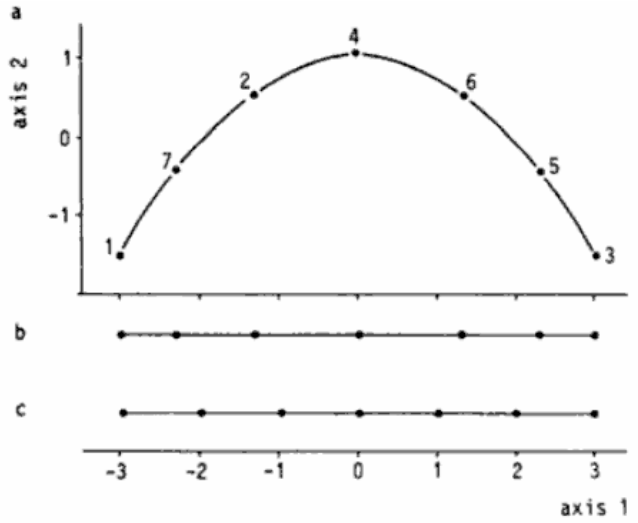
Detrended Correspondence Analysis

The process of "detrending"

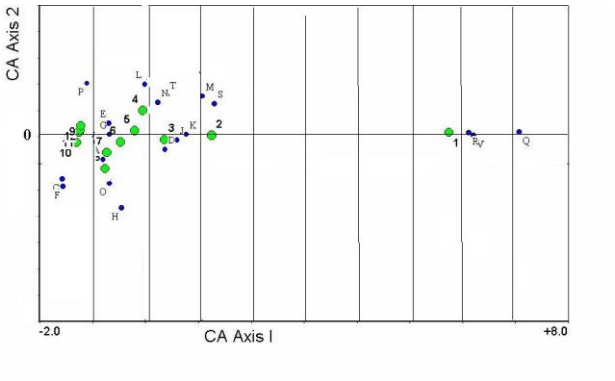
Step 3 – nonlinear rescaling of the first axis, which removes the clumping of samples at the ends of the gradient

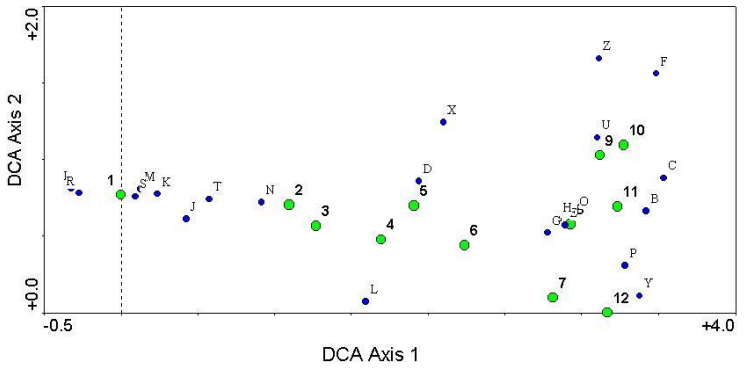
-> resulting ordination diagram has axes scaled in SD values (SD = standard deviation in Gaussian curve)

-> half-change in species composition will occur along gradient of length 1-1.4 SD



ter Braak (1987)





http://ordination.okstate.edu

Detrended Correspondence Analysis

PROS AND CONS

- ☹ inelegant method, which is sometimes compared to the use of hammer on data
- ☹ the result is strongly dependent on the decision about the number of segments (recommendation: do not stick to default only)
- ☹ if there are two or more strong environmental gradients in data, DCA cannot handle them (but this is similar also for other ordination methods)
- ☹ the gradient of the second (and higher) ordination axis is distorted by detrending

- ☺ even hammer, if used by expert, can be an effective tool – the method gives often results with good ecological interpretation
- ☺ axes of DCA are in SD units, allowing for estimation of gradient length

Selection of ordination method

Linear or unimodal?

- linear methods requires homogeneous data, unimodal methods can handle heterogeneous data
- recommendation of Lepš & Šmilauer (2003) – use DCA (detrending by segments) to determine gradient length, and if the first DCA axis is

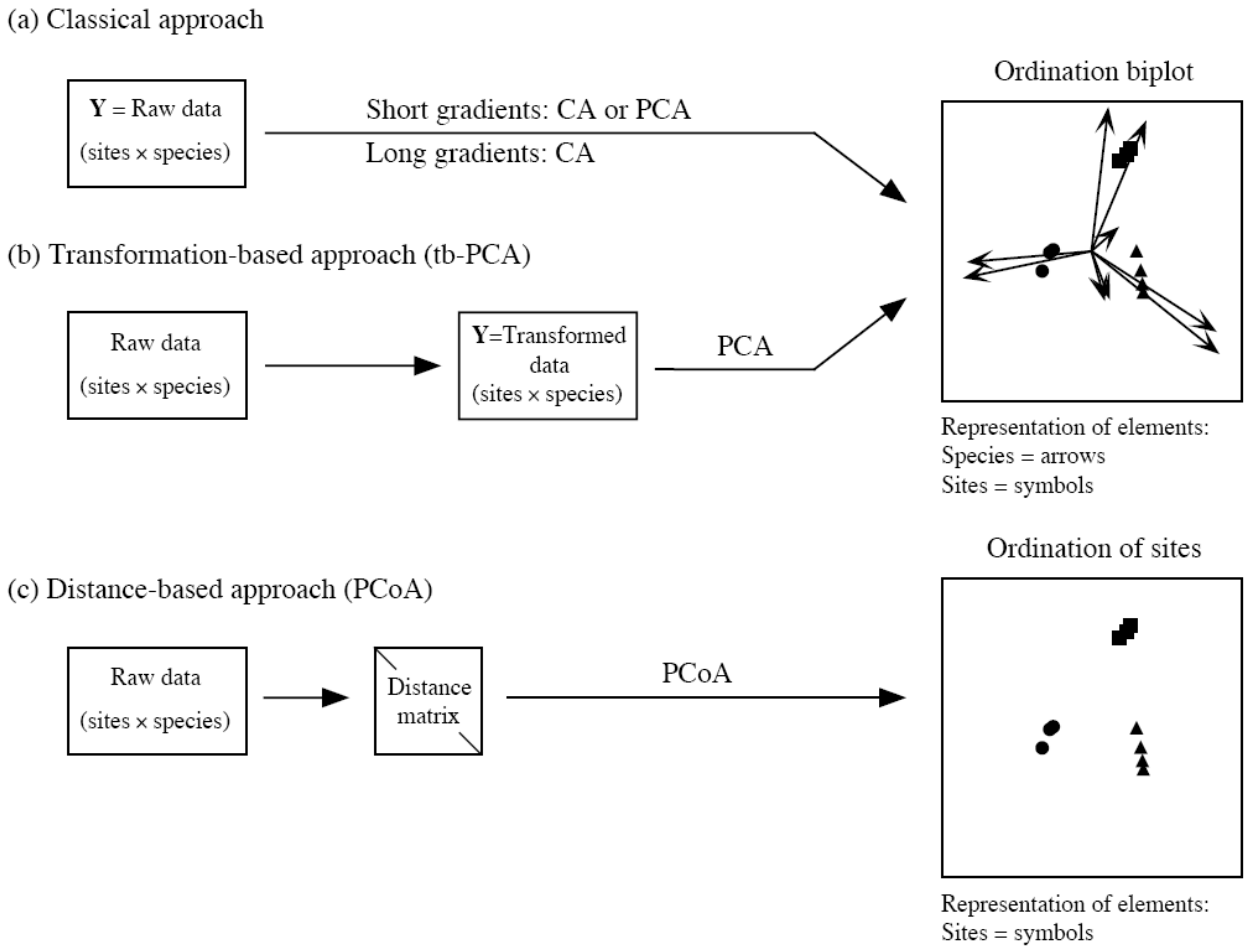
  shorter than 3 SD – use linear technique

  longer than 4 SD – use unimodal technique

  between 3-4 SD – both techniques are OK

- however, this is just a ‘cookbook’ recommendation, not based on research, and doesn’t have to apply in every case

Three approaches to unconstrained ordination analysis



Legendre & Legendre (2012)

PCoA and NMDS

DISTANCE-BASED ORDINATION METHODS

- calculation based on matrix of dissimilarities between samples
- result dependent on the choice of distance measure used

PCoA – Principal Coordinate Analysis (Metric Dimensional Scaling)

- if Euclidean distance is used -> result identical to PCA
- if Chi-square distance is used -> result identical to CA

NMDS – Non-metric Multidimensional Scaling

- non-metric alternative of PCoA
- iterative method, each run can find different solution
- in the beginning, number of dimensions (k) need to be chosen
- with larger datasets VERY time consuming

Comparison of DCA and NMDS

DCA

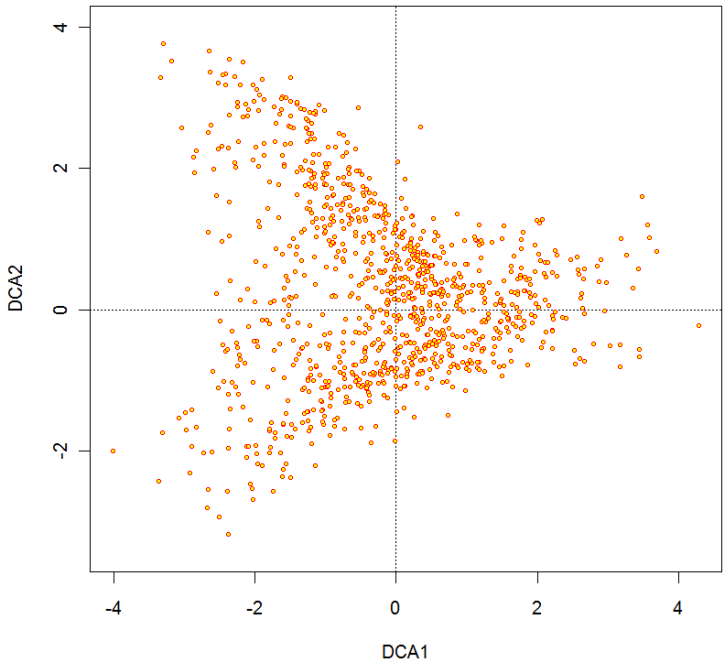
- detrending in reality twists the ordination space, so it looks pretty in 2D, but not so in 3D and higher.
- points will produce triangular or diamond shape – which is actually artifact of detrending!

NMDS

- the method tries to project samples into 2D figure, so as distances between these samples maximally correspond to the sample dissimilarities
- non-metric method – doesn't assume the unimodal shape of species response curves
- according to Minchin (1987) it's the most robust unconstraint ordination method in vegetation ecology

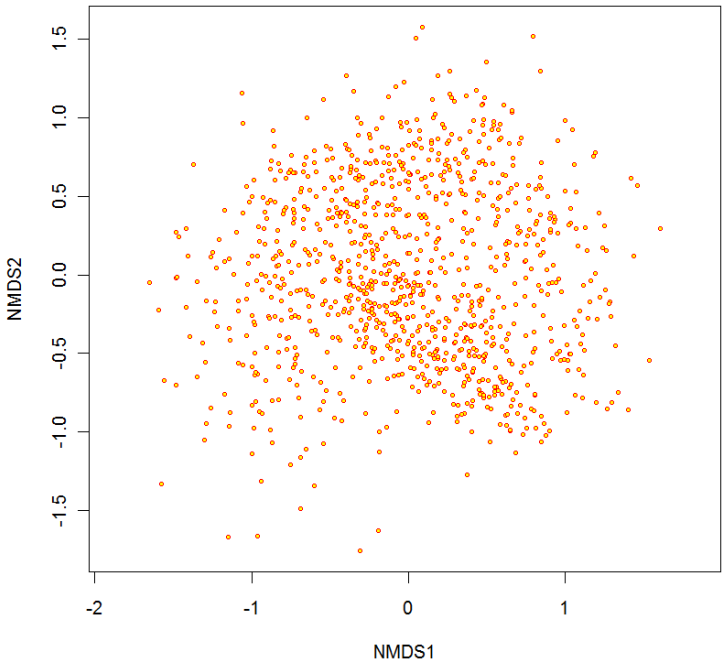
Comparison of DCA and NMDS

DCA



triangle-shape artifact

NMDS



tends to display results as sphere

How to read results of ordination?

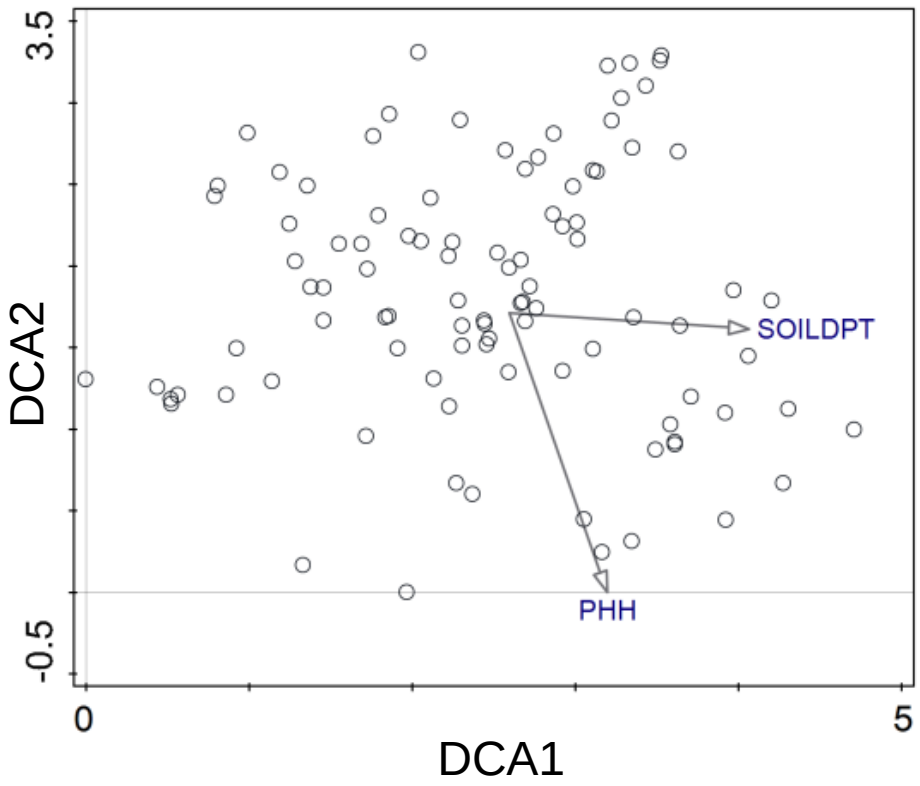
- variability explained by main ordination axes
  - calculated as: **axis eigenvalue / total variance**
  - indicates, how successful the ordination process was
  - the more are species correlated with each other, the more variability will be explained by several main ordination axes
  - **it makes sense to compare** variability explained by various ordination methods on the same dataset
  - **it doesn't make sense to compare** variability explained by the same ordination methods applied on different datasets (*eigenvalues* are dependent on number of players in a game – species and samples)

How to read results of ordination?

- sample scores on ordination axes
  - in ordination diagram samples represented by points (both linear and unimodal techniques)
  - distance between samples in ordination space is proportional to the dissimilarity in their species composition
- scores of independent (environmental) variables *
  - regression coefficients, important are their signs (positive / negative)
- test of significance (Monte-Carlo permutation test) *
  - indicates statistical significance of used environmental variables

* only constrained ordination techniques

Pasively projected environmental variables in uncontrained ordination



species data matrix

	spe1	spe2	spe3	spe4	…
sam1					…
sam2					…
sam3					…
sam4					…
…	…	…			…

DCA

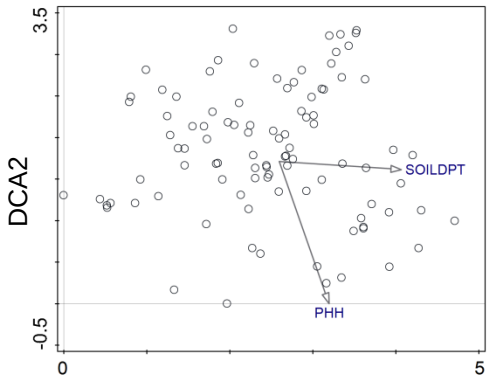
sample scores on the first and second DCA axis

	DCA1	DCA2
sam1		
sam2		
sam3		
sam4		
…	…	…

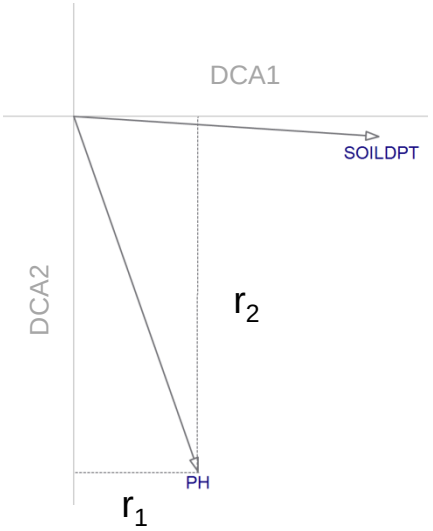
correlation

environmental variables

	PH	SOILDPT
sam1		
sam2		
sam3		
sam4		
…	…	…



ordination diagram DCA



relationship between environmental variables and ordination axes

	PH	SOILDPT
DCA1	r_1	r_3
DCA2	r_2	r_4

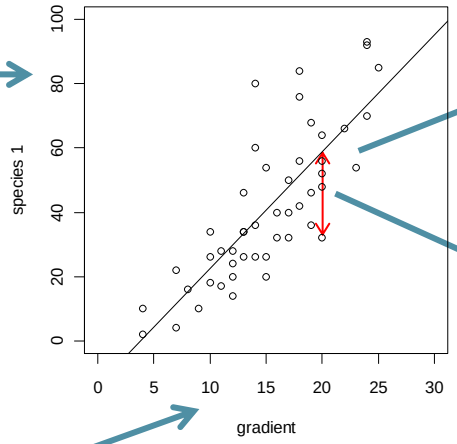
correlation of environmental variables and scores on ordination axes

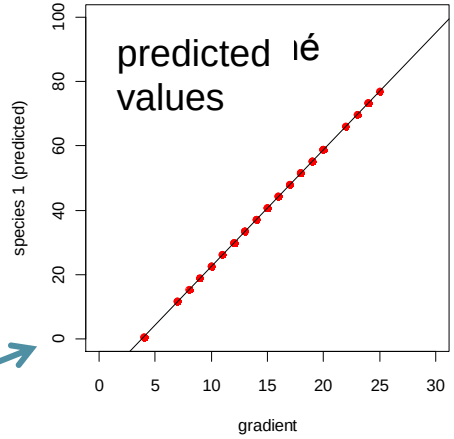
Principle of constrained ordination (RDA)

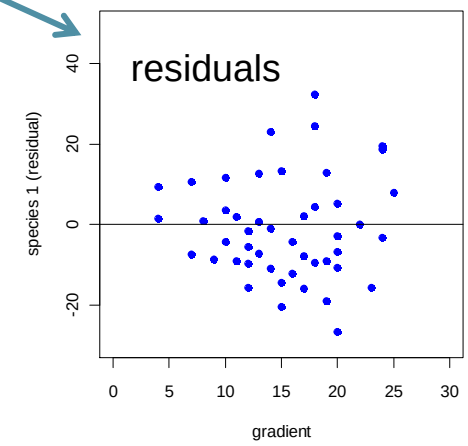
matrix of samples × species

	spe 1	spe 2	spe 3
sam 1			
sam 2			
sam 3			
sam 4			
sam 5			
sam 6			
sam 7			

regression of abundance on environmental gradient







	spe 1	spe 2	spe 3
sam 1			
sam 2			
sam 3			
sam 4			
sam 5			
sam 6			
sam 7			

	spe 1	spe 2	spe 3
sam 1			
sam 2			
sam 3			
sam 4			
sam 5			
sam 6			
sam 7			

	env 1	env 2
sam 1		
sam 2		
sam 3		
sam 4		
sam 5		
sam 6		
sam 7		

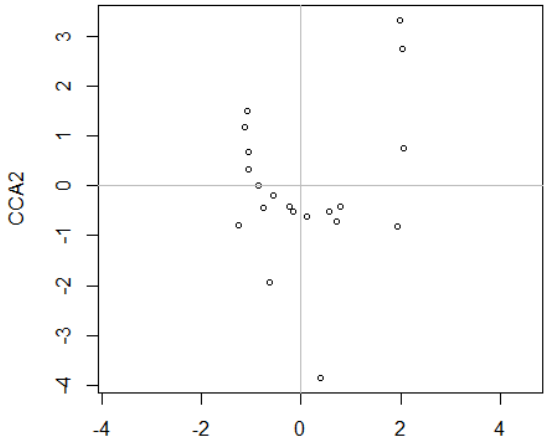
matrix of explanatory variables

matrix of predicted values

	spe 1	spe 2	spe 3
sam 1			
sam 2			
sam 3			
sam 4			
sam 5			
sam 6			
sam 7			

ordinace →

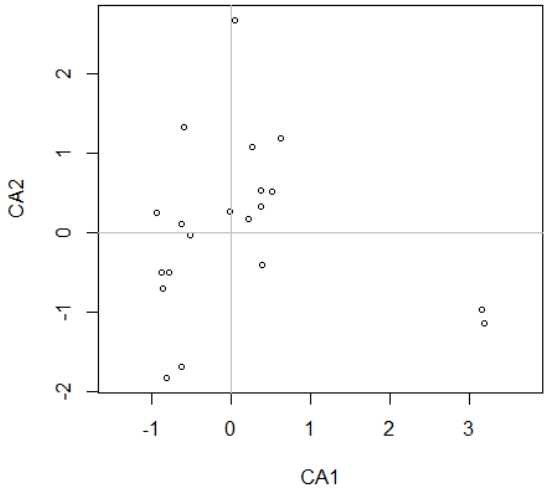
constrained ordination axes



	spe 1	spe 2	spe 3
sam 1			
sam 2			
sam 3			
sam 4			
sam 5			
sam 6			
sam 7			

matrix of residuals

ordinace →



unconstrained ordination axes

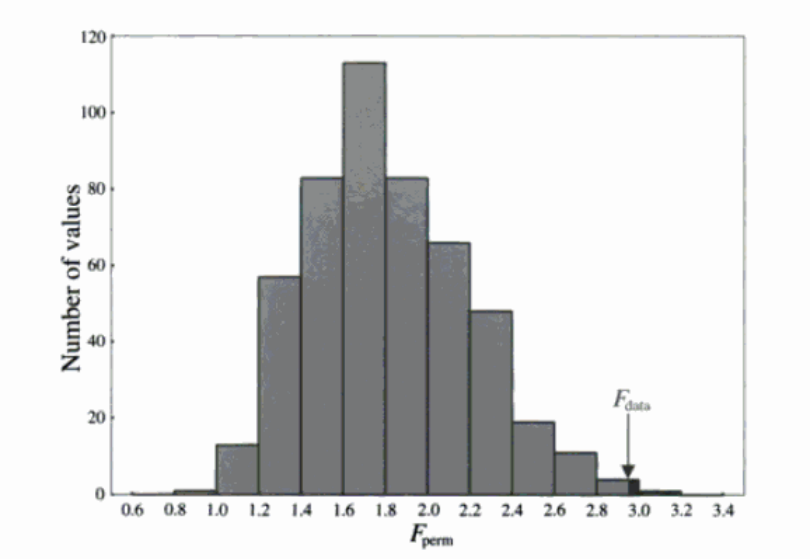
number of constrained ordination axes = number of explanatory varaibles

(if the explanatory variable is categorical, number of constrained axes equals to number of categories -1)

CONSTRAINED ORDINATION ANALYSIS

MONTE-CARLO PERMUTATION TEST

- it tests the null hypothesis, that the species composition is not related to any of environmental variables
- test of the first canonical axis – tests the effect of only one (quantitative) variable
- test of all canonical axes – tests the effect of all environmental variables, or effect of one categorical environmental variable (no of axes = no of categories-1)



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

n_x – no of permutations with $F_{perm} >= F_{data}$

N – no of all permutations

Partial ordination (e.g. pCCA)

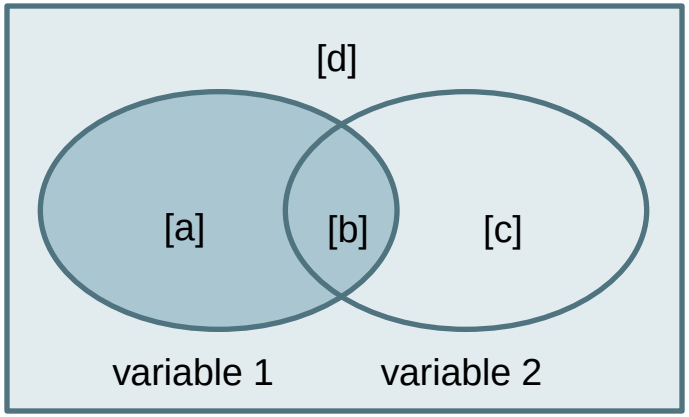
- removes the part of the variability explained by environmental factors, which are not relevant / interesting (e.g. the effect of block)
- „not interesting“ variables are defined as **covariables**
- after this, remaining variability in species data is analyzed using unconstrained or constrained ordination
- if unconstrained analysis follows – ordination axes represent the variability in species composition, which remains after removing the effect of covariables
- if constrained analysis follows – ordination axes represents the net effect of all other environmental variables without the effect of covariables

VARIANCE PARTITIONING

BORCARD ET AL. 1992, *ECOLOGY* 73: 1045–1055

Calculation procedure:

explanatory variable	covariable	explained variability
1 and 2	none	[a]+[b]+[c]
1	2	[a]
2	1	[c]



[a] + [b] – marginal effect of variable 1
[a] – conditional effect of variable 1 (conditioned by variable 2)

shared variance: [b] = ([a]+[b]+[c])-[a]-[c]
variability not explained by the model: [d] = *Total inertia* – ([a]+[b]+[c])

Borcard et al. 1992, *Ecology* 73: 1045–1055

VARIANCE PARTITIONING

ØKLAND (1999) *J. VEG.SCI.* 10: 131-136

- amount of compositional variability extracted by ecologically interpretable ordination axes, if calculated as eigenvalue-to-total inertia ratio, is underestimated due to lack-of-fit of data to model
- the common interpretation of unexplained variability as random variation (noise) in data is inappropriate
- recommendation: do not calculate eigenvalue-to-total inertia-ratio; instead, focus on relative amount of variation explained by different sets of explanatory variables

Forward selection of environmental variables

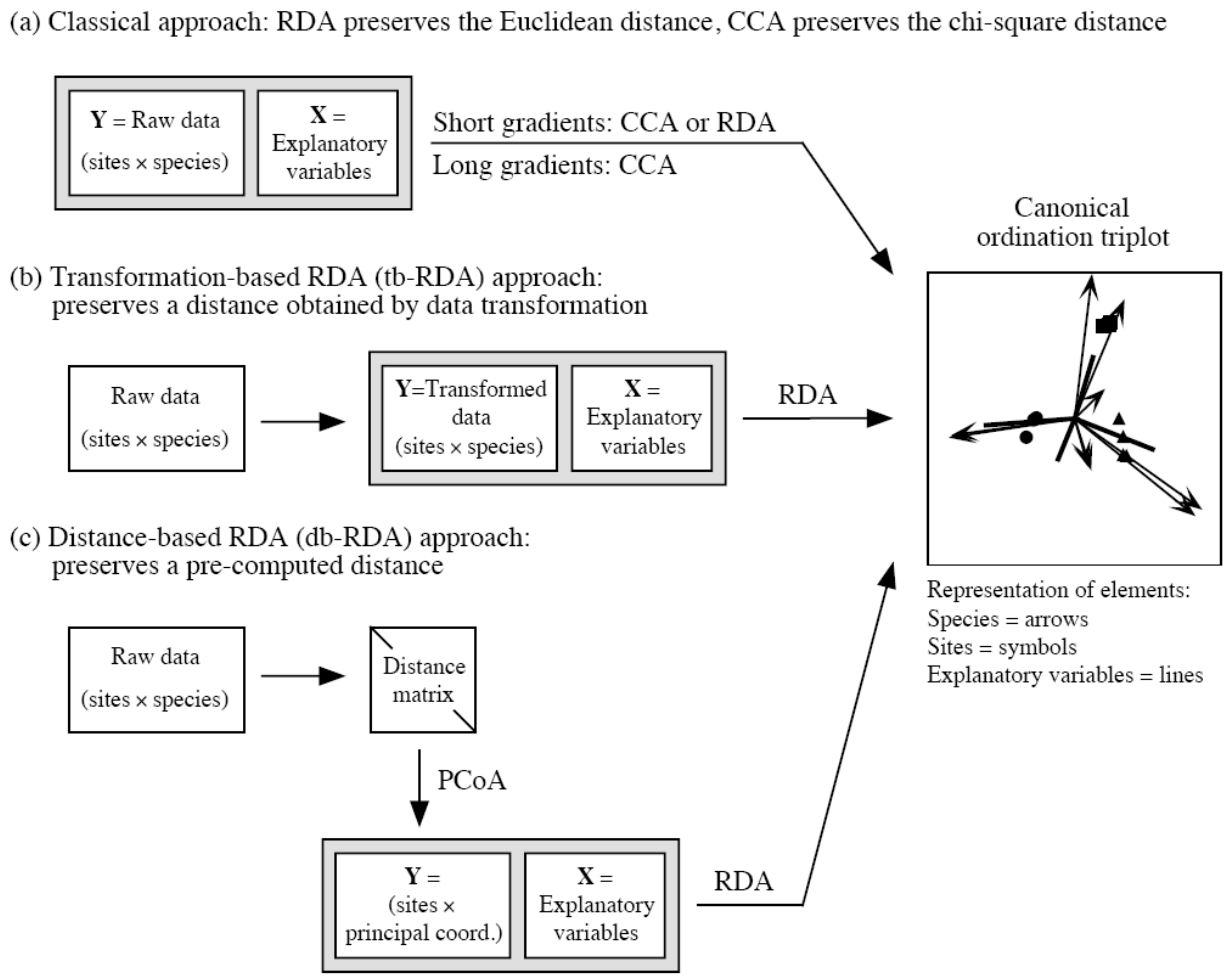
- from available environmental variables choose only those with significant effect
- in each step, significance of particular variables is tested by Monte-Carlo permutation tests
- it selects the variables, which explains the most variability and is significant – this variable is included in the model as covariable
- in the next step, it includes other variables and continues in statistical testing
- significance tests suffer from the problem of multiple comparisons and thus they are quite liberal (number of significant variables included in the model is unrealistically high, Bonferroni correction is desirable)

WHY ORDINATION?

- it is impossible to visualize more than three dimensions – but ecological data have hundreds of dimensions
- reduced low-dimensional ordination space represents main ecological gradients and reduce the noise in data (*ordination* = *noise reduction technique*)
- in case of statistical testing, the ordination doesn't suffer from the problem of multiple comparisons
- we can determine the relative importance of different gradients; this is virtually impossible with univariate techniques
- some techniques (DCA) provide the measure of betadiversity
- the graphical results from most techniques often lead to ready and intuitive interpretations of species-environment relationships

http://ordination.okstate.edu

Three alternative approaches to constrained ordination

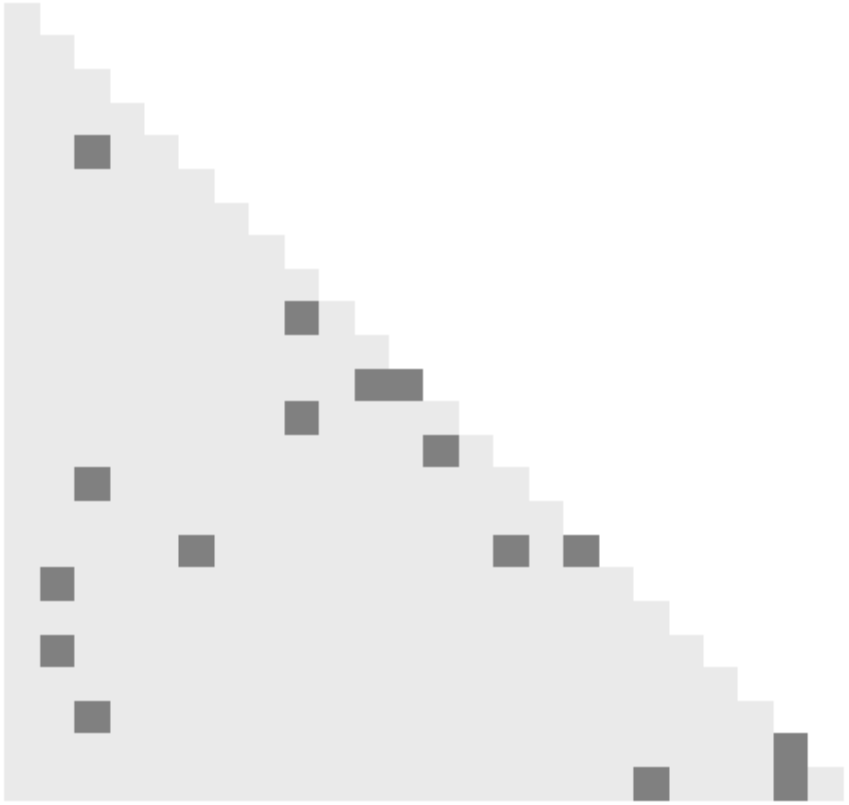


Legendre & Legendre (2012) following Legendre & Gallagher (2001)

Problem of multiple testing

Simulation:

- 25 randomly generated variables
- test the significance of the correlation of each pair
- significant correlations ($p < 0.05$) are represented by dark squares
- total of 300 analyses, 16 significant
- solution: apply correction for multiple testing (e.g. Bonferroni)



MANTEL TEST

environmental variable

	pH
1	4.5
2	4.1
3	4.2
4	3.8

D_e

1	0			
2	0.4	0		
3	0.3	0.1	0	
4	0.7	0.4	0.3	0
	1	2	3	4

species x sample matrix

	sp1	sp2
1	0	3
2	1	2
3	1	2
4	2	1

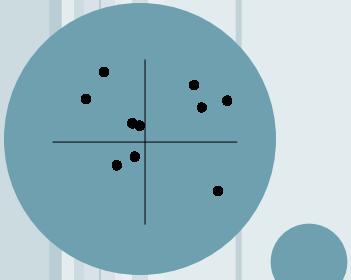
(eucl.)

D_{sp}

1	0			
2	1.41	0		
3	0.3	0.1	0	
4	0.7	0.4	0.3	0
	1	2	3	4

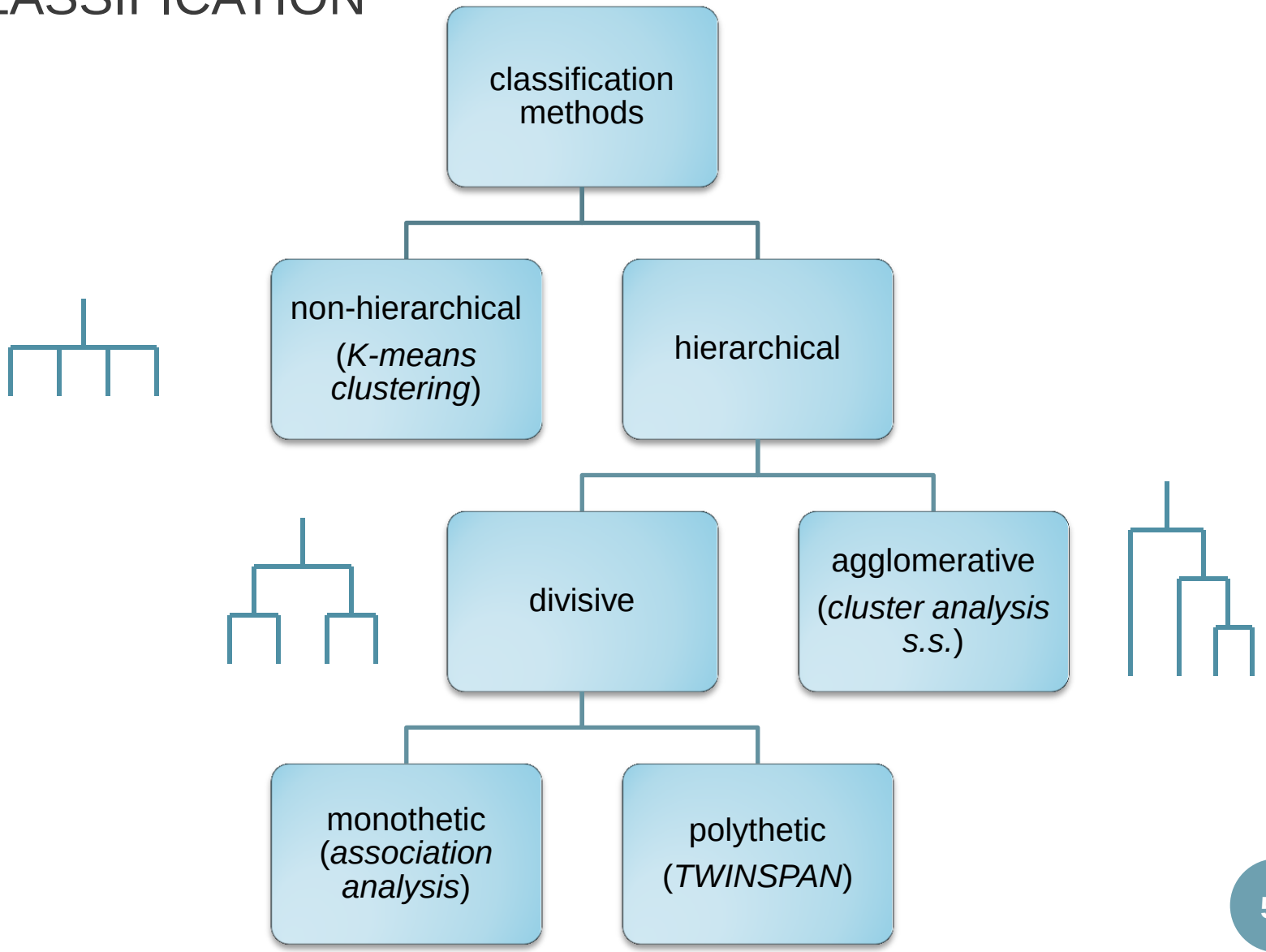
D_e	D_{sp}
0.4	1.41
0.3	1.41
0.1	0
0.7	2.5
0.4	1.41
0.3	1.41

r = 0.965
p = 0.015



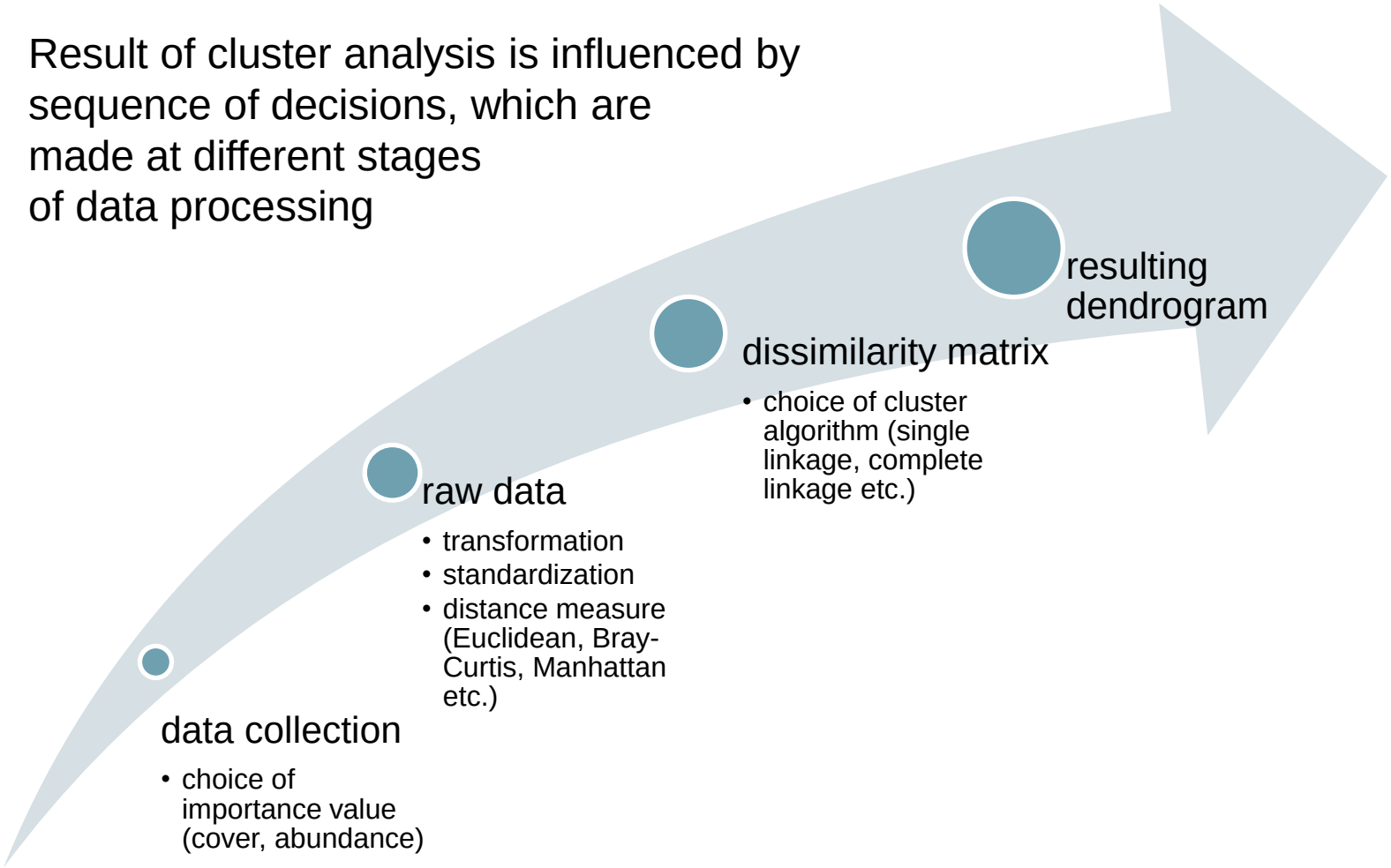
CLASSIFICATION

CLASSIFICATION



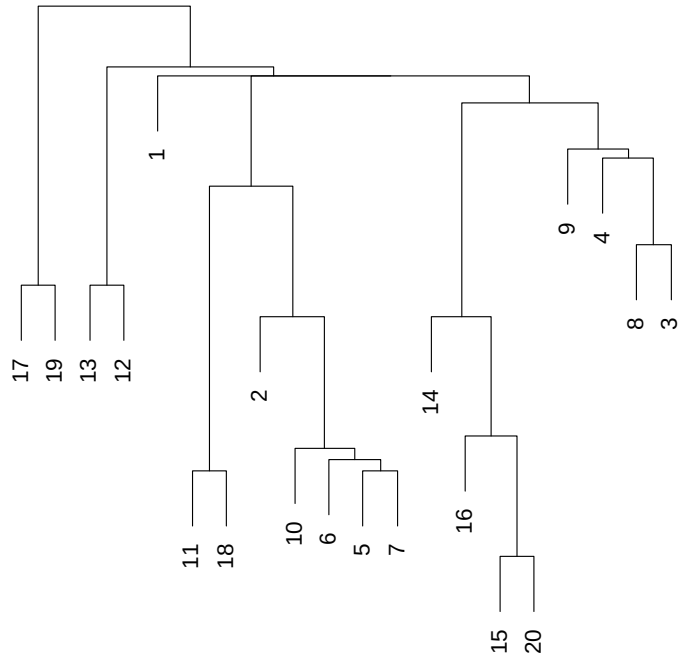
CLUSTER ANALYSIS

Result of cluster analysis is influenced by sequence of decisions, which are made at different stages of data processing

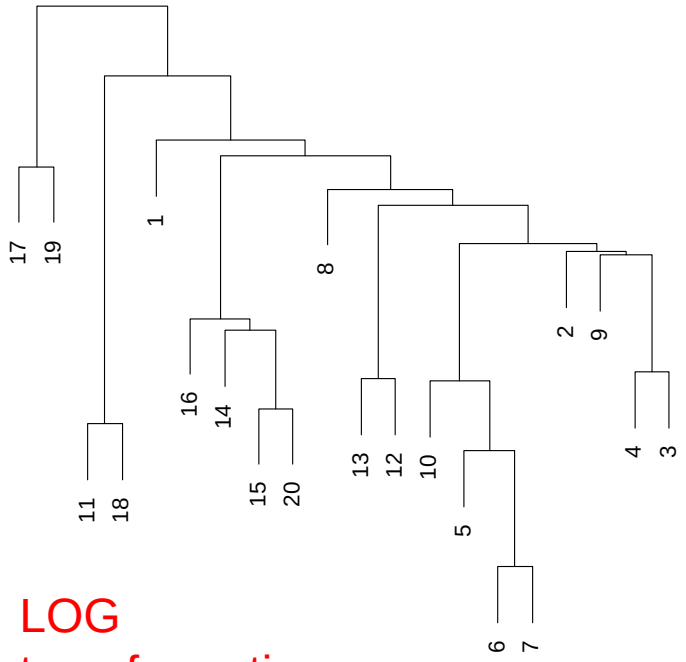


INFLUENCE OF DATA TRANSFORMATION

Single linkage / Euclidean distance / no transformation



Single linkage / Euclidean distance / LOG transformation

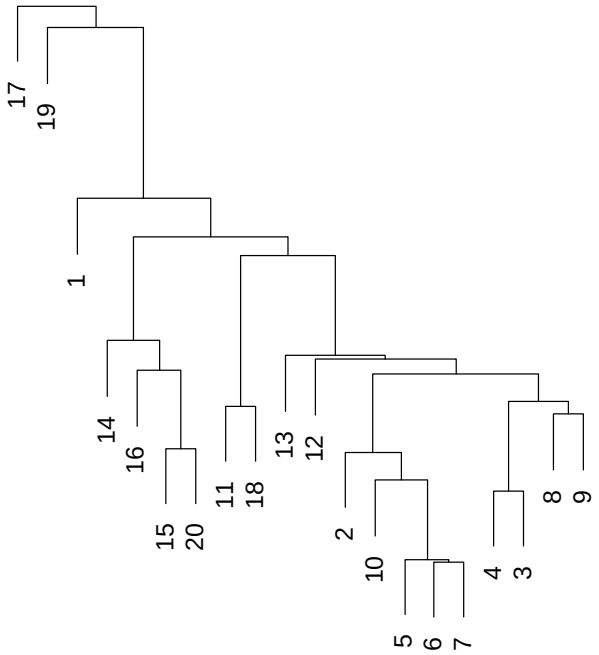


LOG transformation

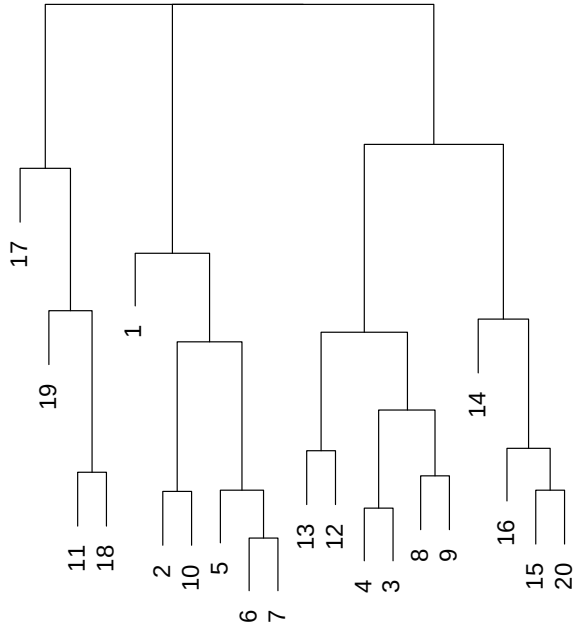
transformation of data (e.g. log transformation) may strongly influence the result of classification (in case of Euclidean distance and single linkage method it is especially true)

SINGLE LINKAGE × COMPLETE LINKAGE

Bray-Curtis distance / Single linkage



Bray-Curtis distance / Complete linkage



single linkage has pronounced chaining

Clustering algorithms

Single linkage (Nearest neighbor)

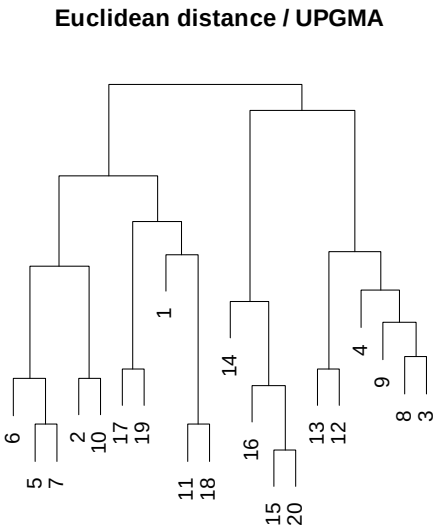
Complete linkage (Furthest neighbor)

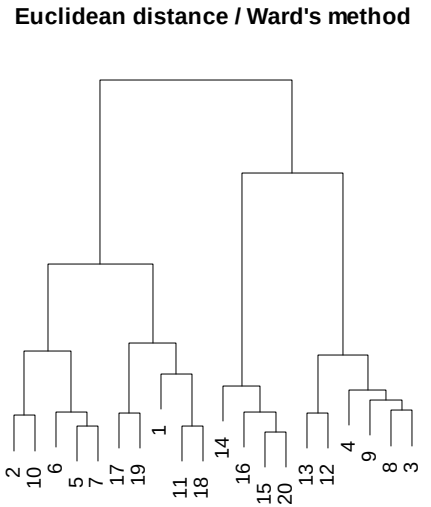
Average linkage

- includes several methods (e.g. UPGMA) standing between single and complete linkage
- more meaningful for ecological data

Ward's minimum variance method

- cannot be combined with semimetric similarity indices (e.g. Sørensen / Bray-Curtis similarity index)

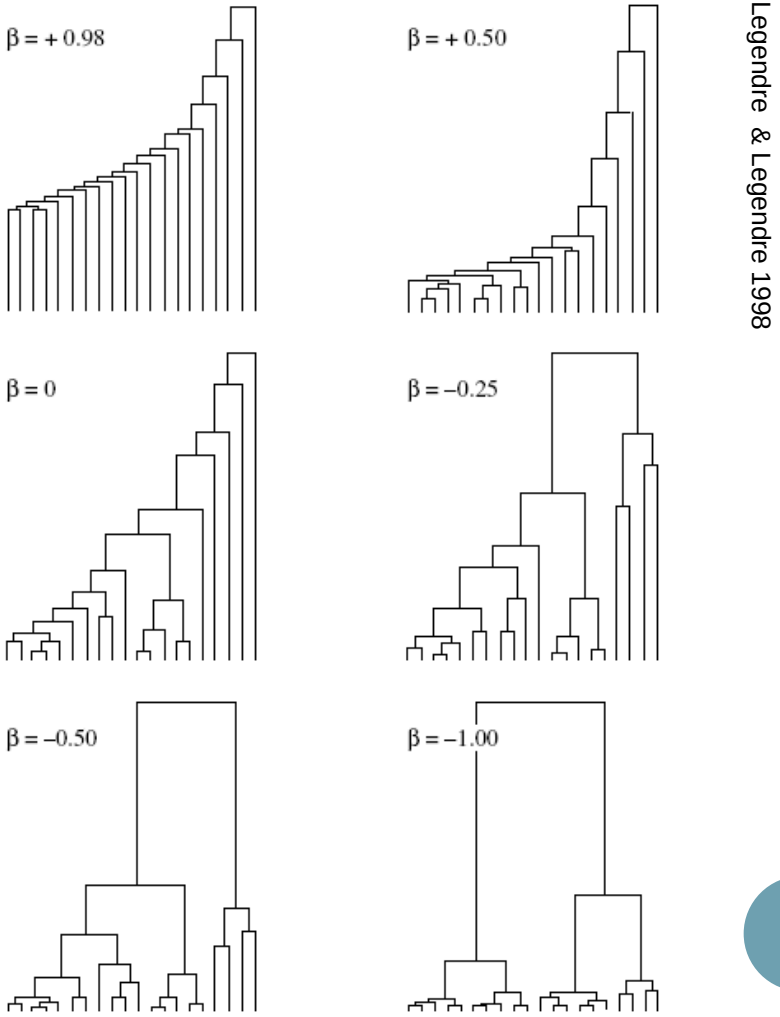




Clustering algorithms

Flexible clustering (beta flexible)

- parameter β influences the chaining of the dendrogram
- highest chaining for β ~ 1, lowest chaining for β = -1
- optimal representation of distances among samples for β = -0.25



Legendre & Legendre 1998

CLUSTER ANALYSIS × TWINSPAN

Cluster analysis

- agglomerative method – clusters are formed from the bottom, by clustering individual samples
- decisions: which distance measure and clustering algorithm to use

TWINSPAN (Two-Way INdicator Species Analysis)

- divisive method – cuts the data from the top
- suitable for data structured by one strong ecological gradient and for determining several (few) groups along this gradient
- results into two-way sorted table, similar to the one used in phytosociology
- algorithm:
  - samples are sorted along the first axis of correspondence analysis (CA, DCA) and then divided into two groups (positive and negative scores)
  - method has complicated way how to treat the samples located close to the axis center, which have high probability of being misclassified
- includes number of arbitrary numerical steps (often criticized, but still favorite)
- decisions: stopping rules for division, pseudospecies

CLASSIFICATION IN GENERAL

Subjective

- based on subjective decisions of the researcher, not easy to be reproduced by somebody else

Formalized (not objective!)

- selection of clearly defined classification criteria, easy to reproduce

- unsupervised
  - numerical methods of classification (e.g. *cluster analysis*, TWINSPAN)
  - allow only for very rough control of the classification result (you can choose the method and set up several parameters)
- supervised
  - ANN – artificial neural networks, classification trees, random forests, COCKTAIL
  - need to be trained first and than the method can reproduce the same classification structure