Multivariate statistics in R

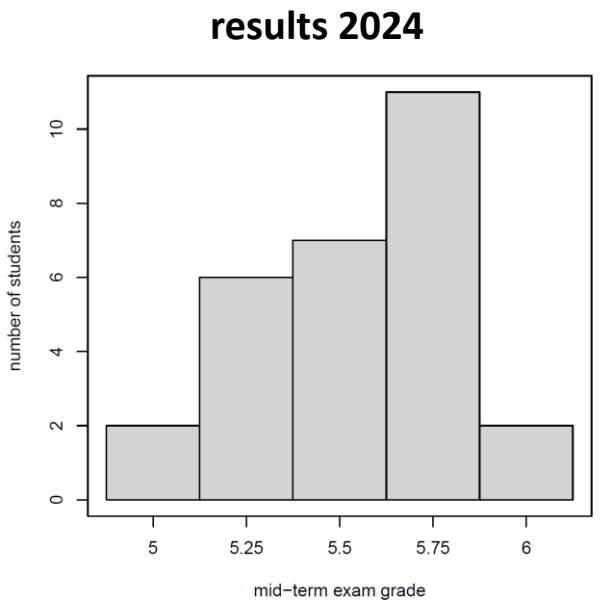
Hannes PETER
Martin BOUTROUX
Zhe LIU

Updated schedule

- 10.09. **session 1**
- 17.09. **session 2**
- 24.09. **group work**
- 01.10. ***«modern R» with Martin*** (tidyverse)
- 08.10. **session 3**
- 15.10. **session 4**
- 22.10. autumn holidays
- 29.10. **group work**
- 05.11. **session 5**
- 12.11. **mid-term exam**, **group work**
- 19.11. **session 6**
- 26.11. **session 7**
- 03.12./10.12./17.12. **group work**
- 07.01. group presentations 1
- 14.01. group presentations 2

mid-term exam

- multiple-choice exam
- app. 10 questions
- general understanding
- no formula
- no R functions



Example:
The double-zero problem in multivariate statistics concerning ecological data refers to…:

a) the inability of Ecologists to measure close to detection limits (hence leading to many zeros)
b) the issue that shared absences of species across sites are typically not ecologically meaningful and should not contribute to similarity
c) that species with absences in two sites can not be modeled using Euclidean distances
d) that double zeros are non-integer values

Ordination

What is the principle of ordination?

-example: PCA

How to interpret the results of an ordination?

- explained variance
- biplots

What are (some of) the different techniques available for unconstrained ordination?

- CA, NMDS, PCoA

a step back…

k-means partitioning

- split dataset into a pre-determined number of groups
- repeated a large number of times (*nstart*) using different random initial configurations until best solution is found (smallest SSE)

the problem: applying k-means to distance matrix

- k-means is not appropriate for species abundance data with lots of zeros
- one possibility is to use non-Euclidean dissimilarity matrices, like Bray-Curtis
- such matrices should be square-root transformed and submitted to principal coordinate analysis in order to obtain a full representation of the objects in Euclidean space. The resulting PCoA axes can then be subjected to k-means partitioning.
- Another solution is to pre-transform the species data (e.g. Chord transformation). Distances are then Chord distances.

Overview of multivariate tools

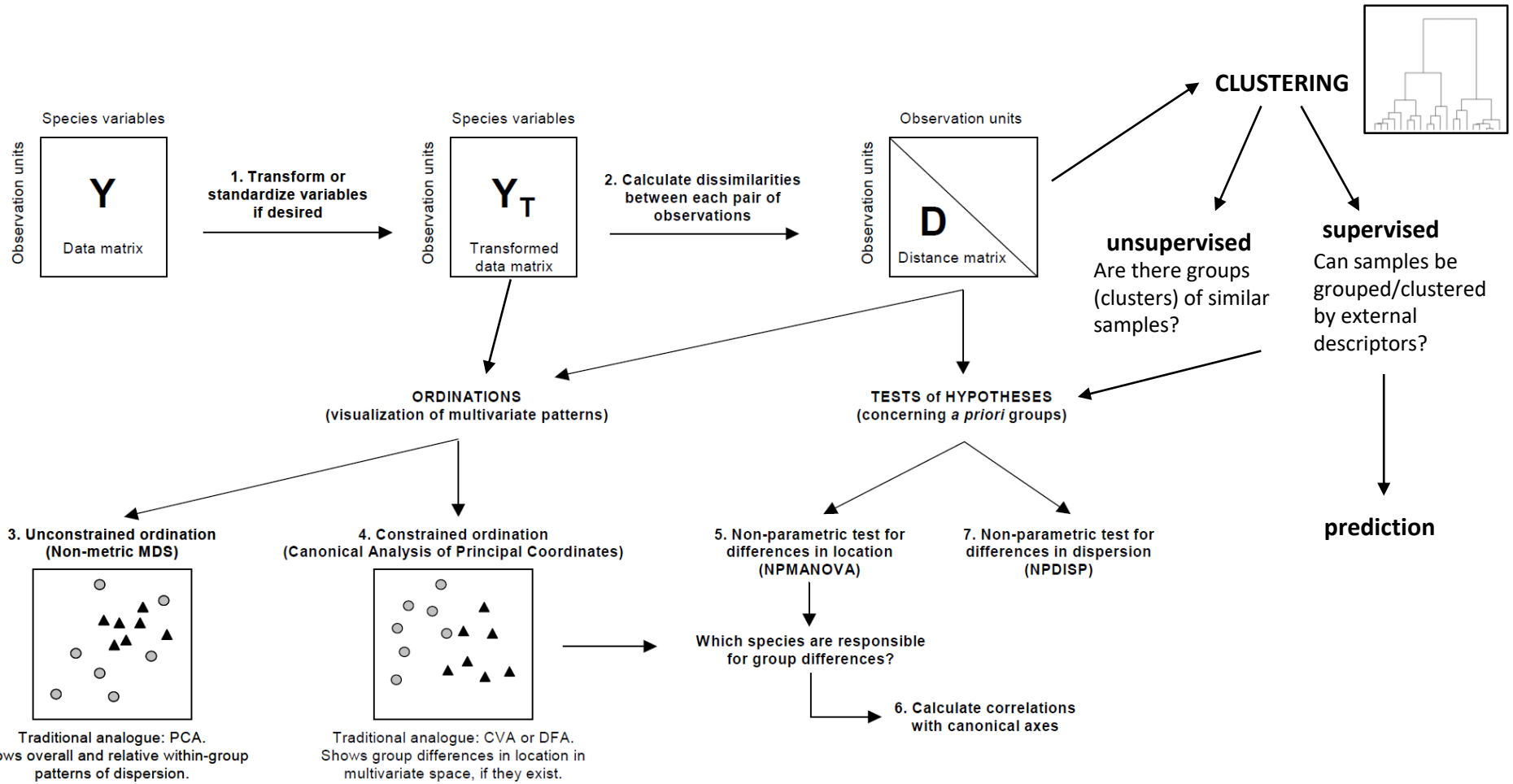
- Similarity (Resemblance)

- Clustering → *searches discontinuities, focus on pairwise ("fine") relationships*
  - Unsupervised
  - Supervised

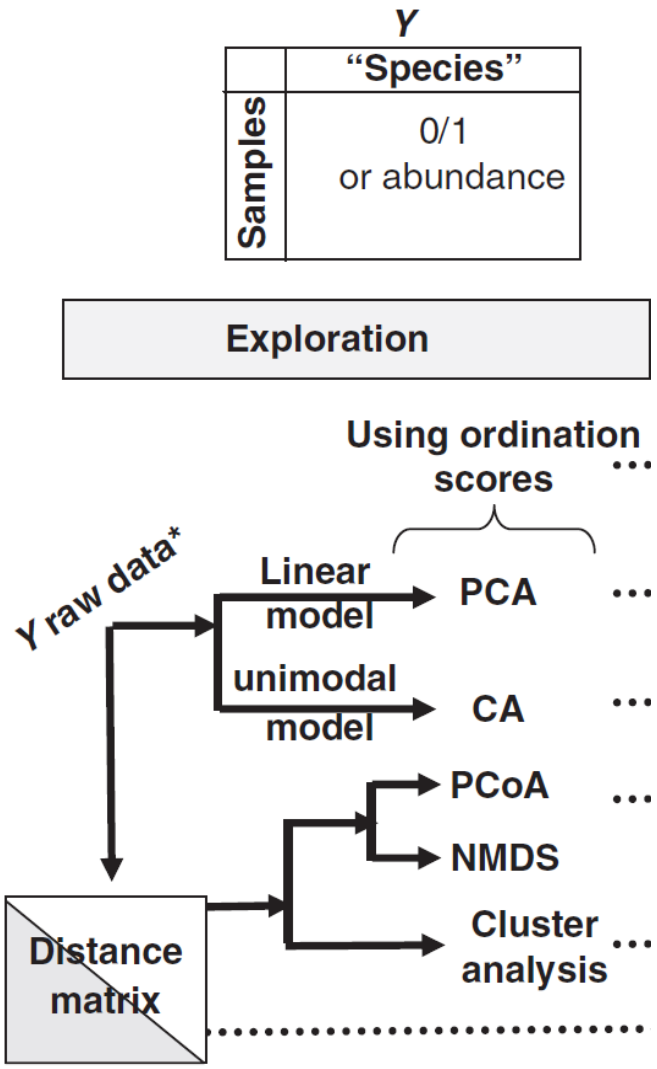
- Ordination (Latin: *ordinatio* setting in order)
  - Unconstrained → *searches main trends (general gradients)*
  - Constrained

complementary

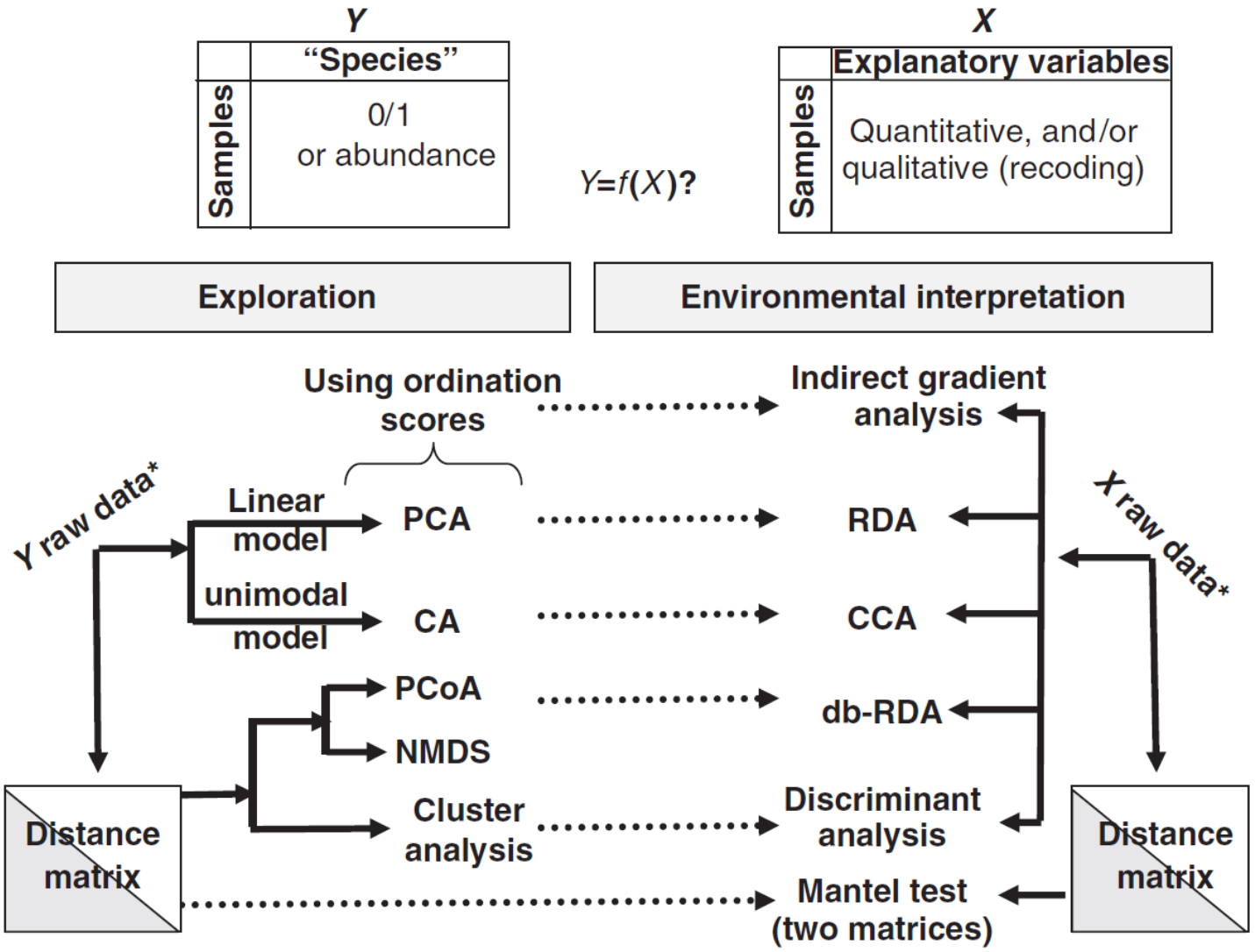
Overview of multivariate tools



even more overview...



even more overview…



Goal of ordination

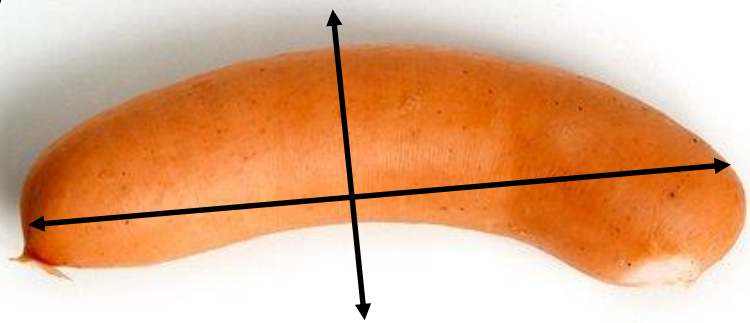
- **Reduce dimensionality**
  - reduce the number of dimensions for visualization typically 2 dimensions
  - identify groups of similar samples
    - ⇒ to simplify the multivariate dataset (hypothesis generation)

Additionally:

- identify taxa that drive (dis)similarities between samples (species scores)
- identify environmental variables that explain the (dis)similarities in taxa composition between samples

Principles of unconstrained ordination

- identify principal axes (dimensions/components/factors)
  - first axis explains most of the differences among samples
  - each axis is independent



- project points (objects) onto axes

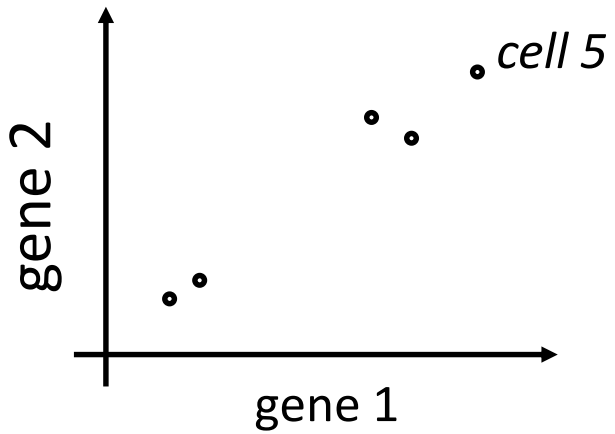
- maintain as much of the variance of the dataset as possible
  - how many components should be shown?

Principal Component Analysis (PCA)

- Archetype of ordination methods
- Descriptors should be quantitative (standardized)
- PCA performs a rotation and a translation of the principal axes of a dispersion matrix (covariance and correlation).
- PCA maintains the distances (Euclidean) between objects in reduced space
  - Usually inadequate for community data (double zero problem) but good for environmental data

Principals of Principal Component Analysis (using Singular Value Decomposition, SVD)

	Cell 1	Cell 2	Cell 3	Cell 4	Cell 5
Gene 1	5	4	8	9	11
Gene 2	7	3	9	8	12

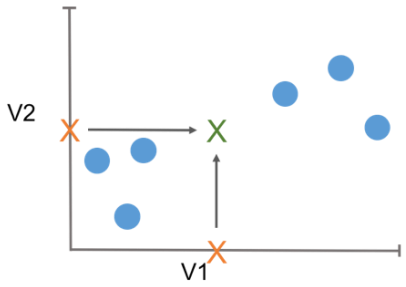


find center

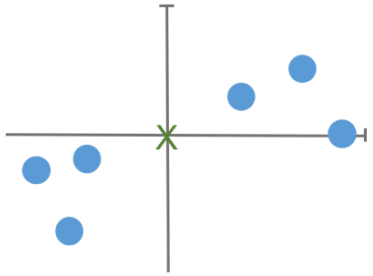
center

line through origin

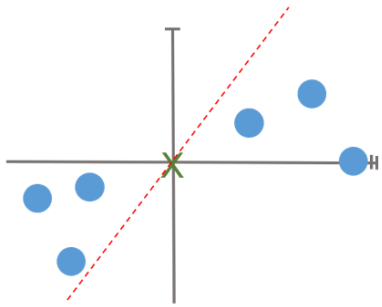
distance of projection to origin



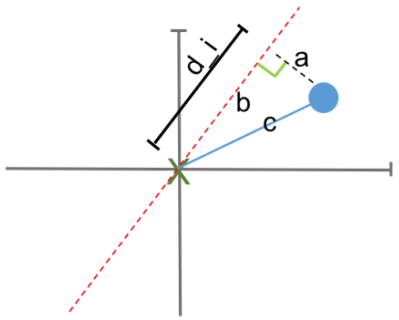
(1)



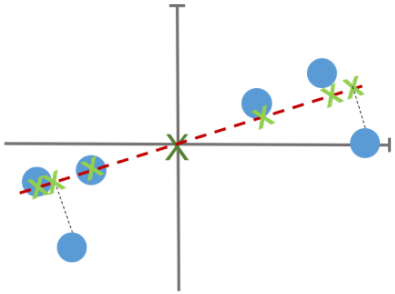
(2)



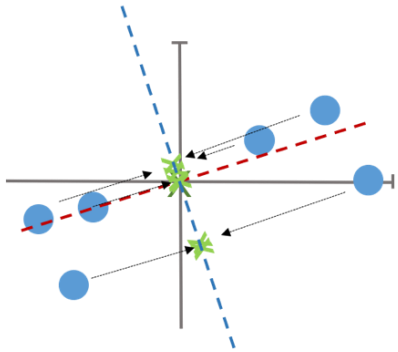
(3)



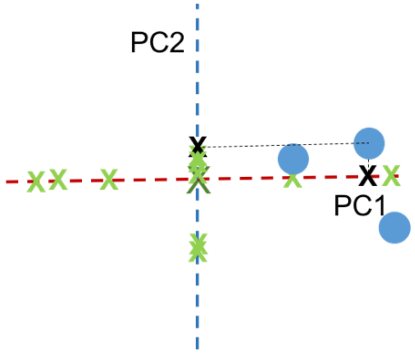
(4)



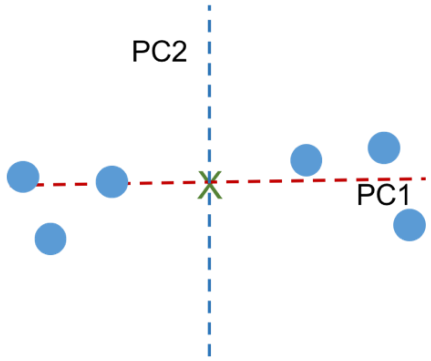
(5)



(6)



(7)



(8)

find main gradient; maximize sum of squared distances of projections to origin (rotation)

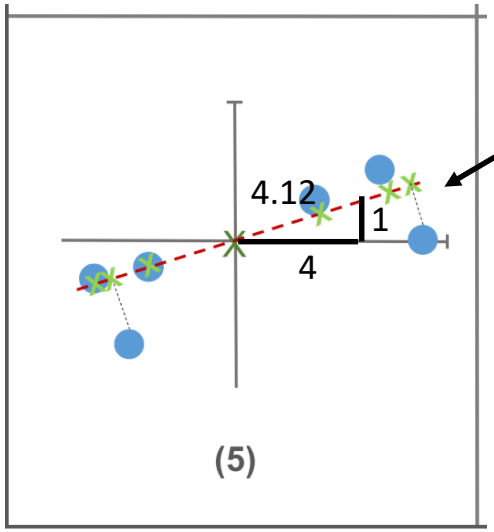
add orthogonal axis through origin; max. SS(dist) for second PC

rotate axes, scale (unit vector)

continue with fitting 3rd PC orthogonal to 2nd PC, etc…

EPFL

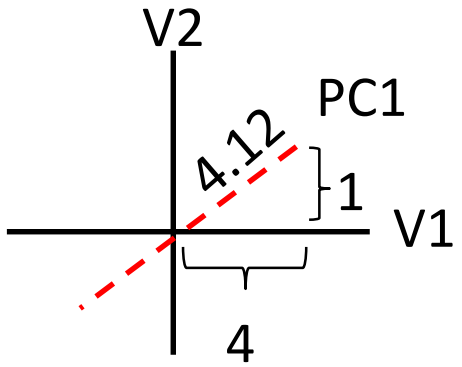
scaling



Principal Component (PC) 1

for example:
slope of PC1: 0.25
4 units on PC1, 1 unit on PC2

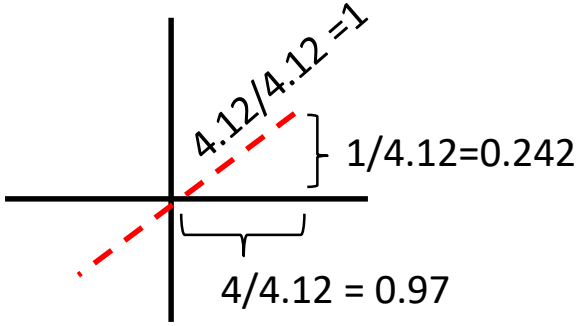
unit vector = Eigenvector

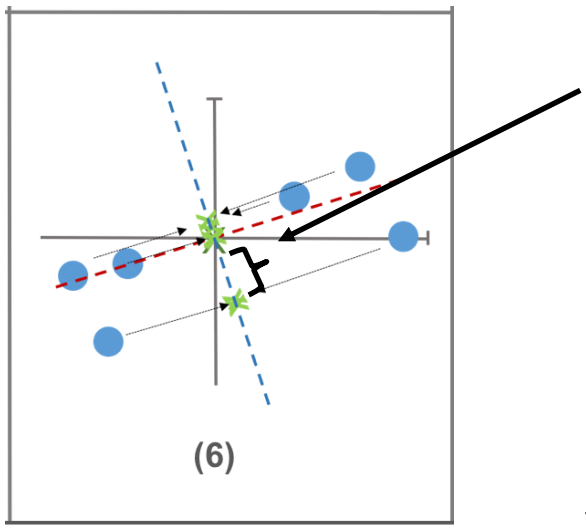


Eigenvector is used for scaling
(e.g. devided by 4.12)

$$\frac{SS(distances\ PC1)}{n-1}$$ = Eigenvalue for PC1

Variation explained by PC1

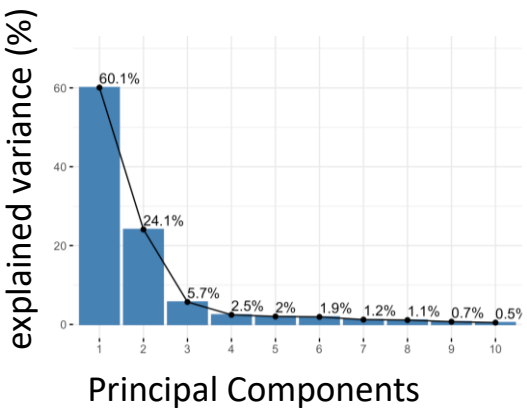


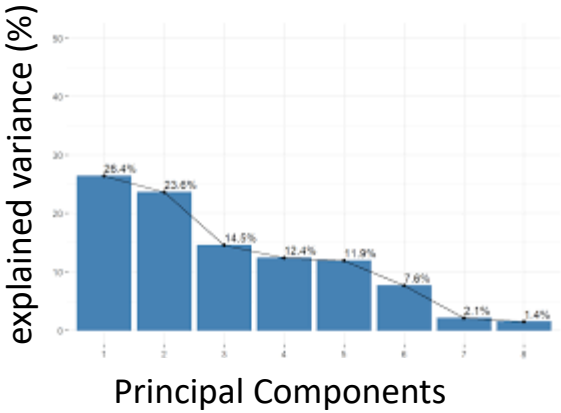


Sum of squared (SS) distances on Eigenvector (here PC2) is called «Eigenvalue of PC2»

$$\frac{SS(dist)}{n-1} = \text{variance (of PC2)}$$

variance of all PCs
=> used in scree plots





Selection of the dispersion matrix (S)

covariance

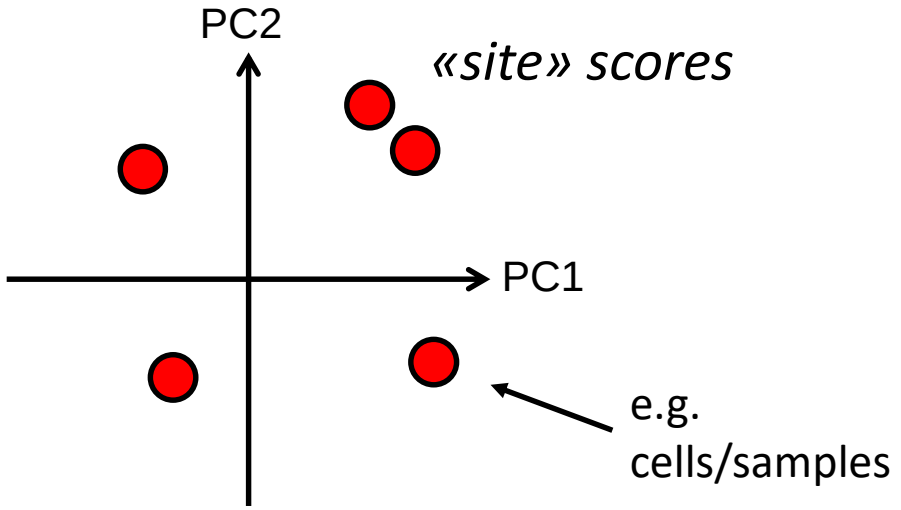
- can be used when descriptors are similar or the same units (or are transformed prior to PCA)
- preserves the variance of descriptors

correlation

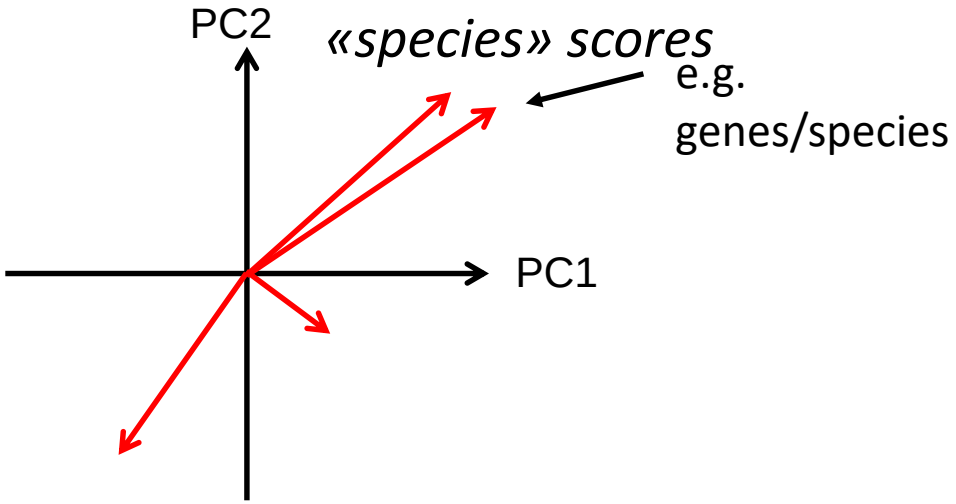
- when descriptors have different units
- removes the variance of the descriptors, thus giving all of them the same weight
- default in vegan::rda

VISUALIZATION

Scatter plot with projection of objects (point symbols)



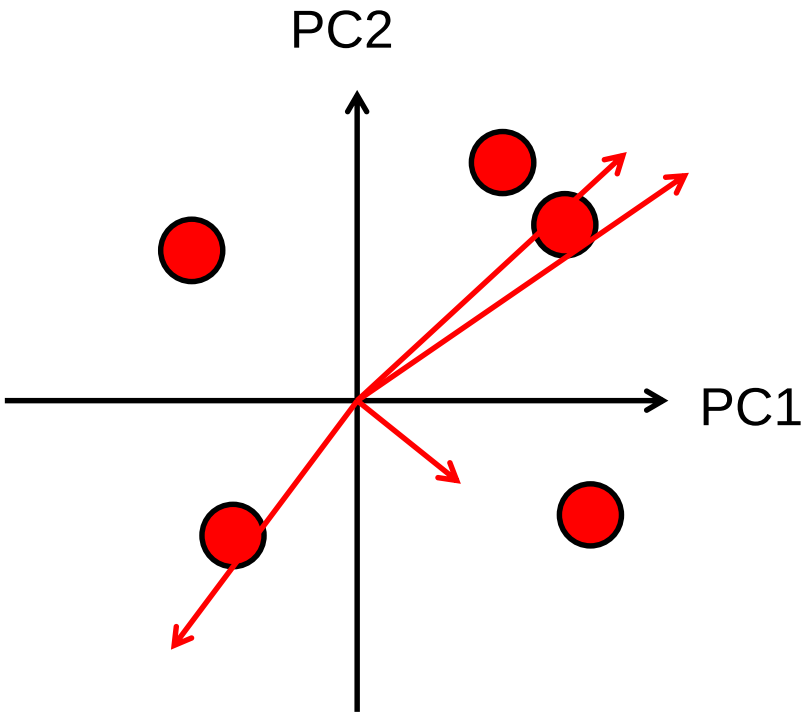
Scatter plot with projection of descriptors (arrows)



arrow length reflects importance of descriptors!

Double projection (*biplot*)

- Shows both objects (symbols) and descriptors (arrows) on the same ordination diagram.

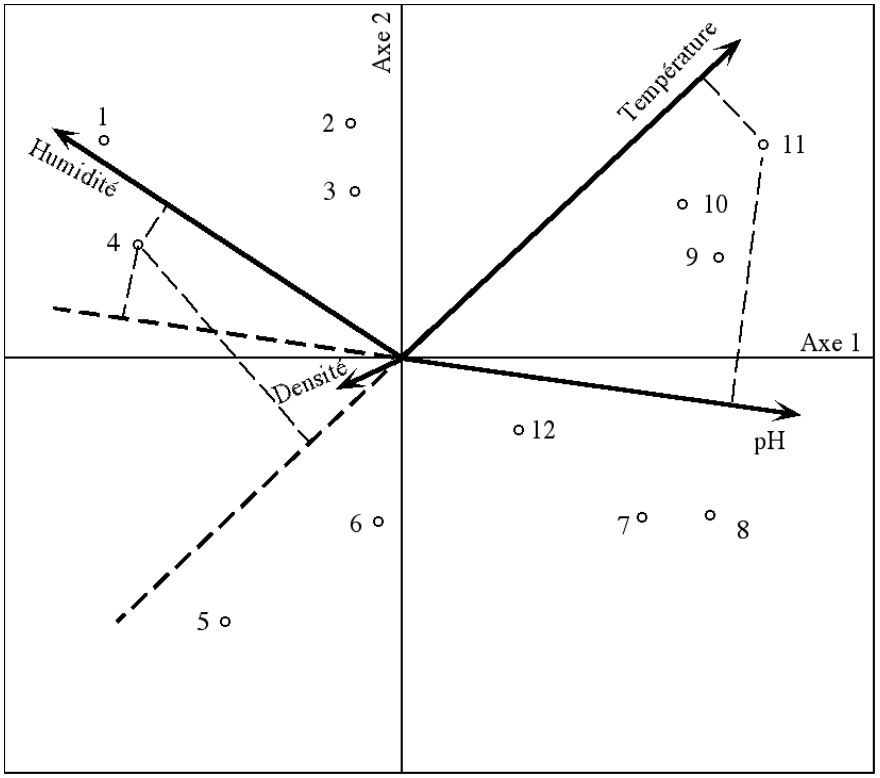


Double projection (*biplot*)

- Two types of scaling are available depending on the purpose of the projection:
  - *Scaling 1* : Focus on species (Norm of eigenvectors = 1)
  - *Scaling 2* : Focus on sites (Norm of eigenvectors = square root of eigenvalues)
  - *Scaling 3* : A compromise

PCA interpretation

PCA on 4 physicochemical soil variables (pH, temperature, humidity, density) measured on 12 sites (1-12)

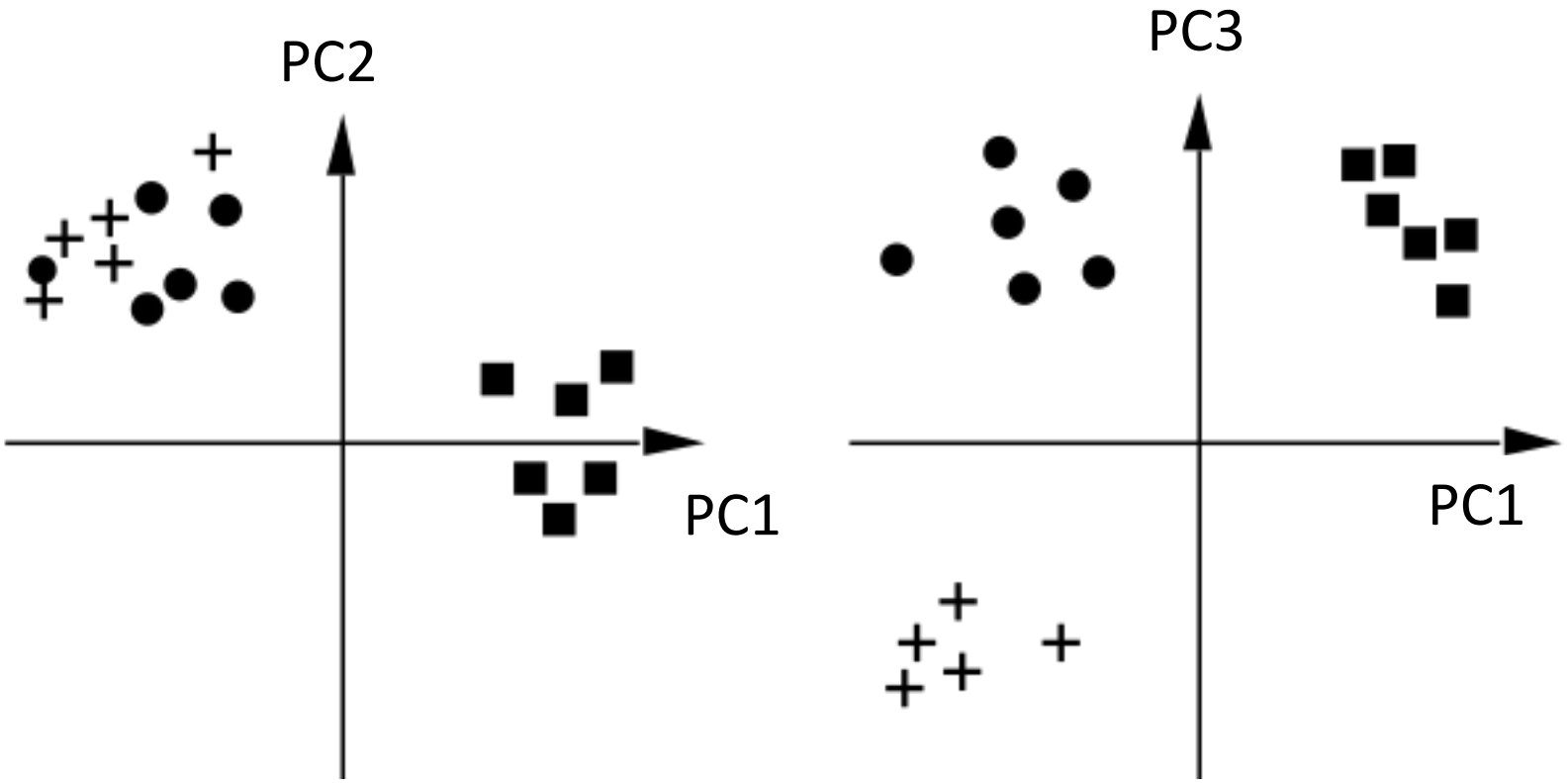


- The proximity between sites 9, 10 and 11 indicates that they have similar soil features (high resemblance)
- Site 11 has a high pH (relative to the other sites), high temperature but low humidity
- *Density* contributes little to explaining differences among the 12 soil types (short arrow)

considerations

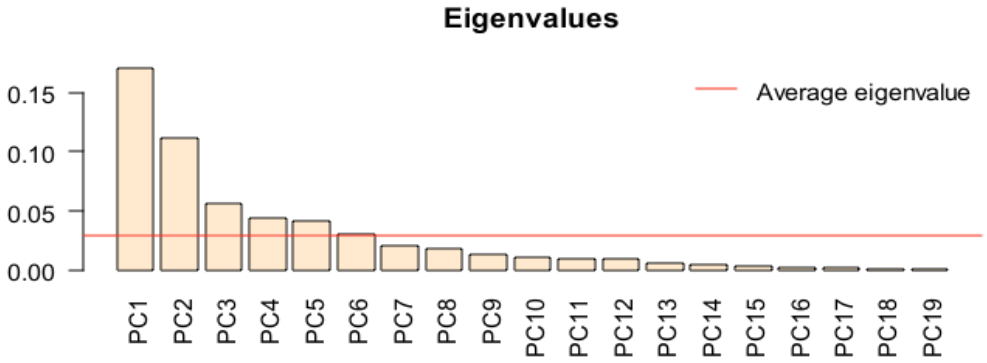
- PC 1 and 2 indicates two groups of objects
- PC 1 and 3 shows three distinct groups of samples
- could also be demonstrated by superposing clustering results (e.g. symbols)

=> plot several PCs (1 and 2, 1 and 3, 2 and 3, ...) before interpreting the proximities of objects

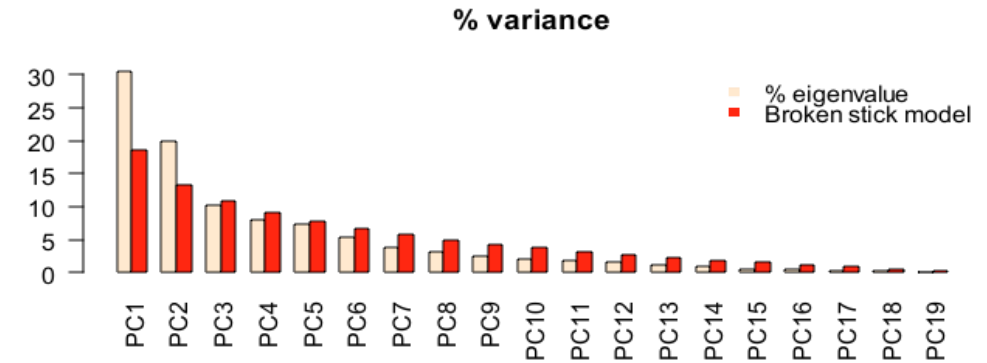


Number of PCs - scree plots

- General idea of PCA: reduce dimensions - showing more than 3 PCs is typically not useful
- Kaiser's rule: axes which contribute to more than average of explained variance
- Broken stick model: principal axes which explain less variance than a random model (broken stick) should not be interpreted



Kaiser's rule:
5 PCs



Broken stick model:
2 PCs

Apply PCA to species abundance data…?

- PCA preserves Euclidean distances and therefore considers the double absence of a species as resemblance (double zero problem)
  - use Hellinger transformation (see PCoA).
- The underlying model assumes a linear response of the descriptors the principal axes
  - This assumption is valid only if the gradients are short
- Alternative ordination techniques for species count data…

Overview of different ordination techniques

Method	Distance preserved	Variables
Principal component analysis (PCA)	Euclidean distance	Quantitative data, linear relationships (beware of double-zeros)
Correspondence analysis (CA)	χ^2 distance	Non-negative, dimensionally homogeneous quantitative or binary data; species frequencies or presence/absence data
Principal coordinate analysis (PCoA), metric (multidimensional) scaling, classical scaling	Any distance measure	Quantitative, semiquantitative, qualitative, or mixed
Nonmetric multidimensional scaling (nMDS)	Any distance measure	Quantitative, semiquantitative, qualitative, or mixed

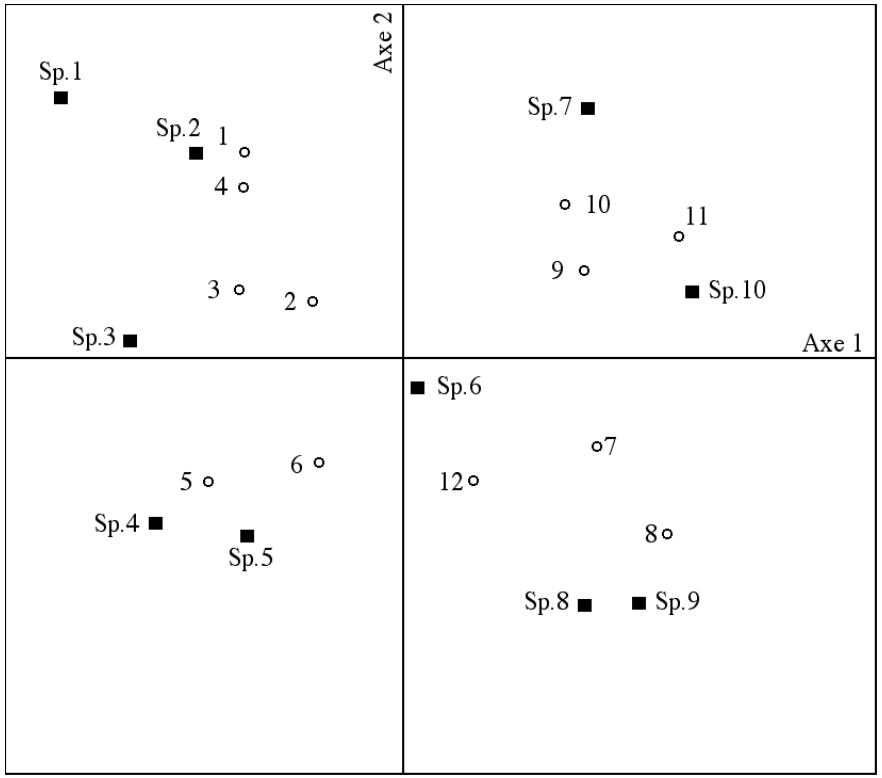
Legendre & Legendre 2012

Correspondence Analysis (CA)
also: Reciprocal Averaging

- Represents the correspondence between the rows and columns of a double centered table of frequency
  - Corresponds to a PCA on a covariance matrix of conditional probabilities
  - Preserves Chi-square distances
  - Ignores double zeros
- Suitable method for presence-absence or abundance data
  - BUT: CA is sensitive to infrequent species
  - Long gradients generate horseshoe effects
- Descriptors should be of the same nature and expressed in the same unit
- Negative values and null vectors (objects or descriptors) in the data matrix are not supported

CA interpretation

CA of an abundance table of 10 species observed at 12 sites



- Sites 1, 2, 3, and 4 have a similar composition (relative frequencies of species)
- Species 1, 2 and 3 are abundant in these sites
- Species 8 and 9 are absent or very sparse in these sites
- Species 6 is present (or absent) at almost all sites

Principal Coordinate Analysis (PCoA)

- Uses any resemblance matrix (similarity matrix)
  - PCoA preserves distance or similarity of the selected association measure.
  - proximity of objects indicates resemblance

- Allows the use of all types of variables (qualitative, semi-quantitative, quantitative), and even to mix them, provided an adequate association measure is chosen (e.g. Gower)

- Doesn't allow the joint analysis of objects and descriptors (such as in CA or PCA; biplots)
  - either a scatter plot of the objects (Q mode) or of the descriptors (R mode).
  - possibility to project descriptors *a posteriori* onto PCoA

Application of PCoA to data with species abundances

- PCoA can produce negative eigenvalues.
  - these axes are not interpretable.
  - choice of distance measure important
- Objects are often better dispersed as compared to CA (no agglutination)
  - less sensible to rare species
- Can be combined with a cluster analysis obtained with the same distance matrix
  - Projection of objects onto the ordination
  - Superposition of the dendrogram.

Non-Metric Multi-Dimensional Scaling (NMDS)

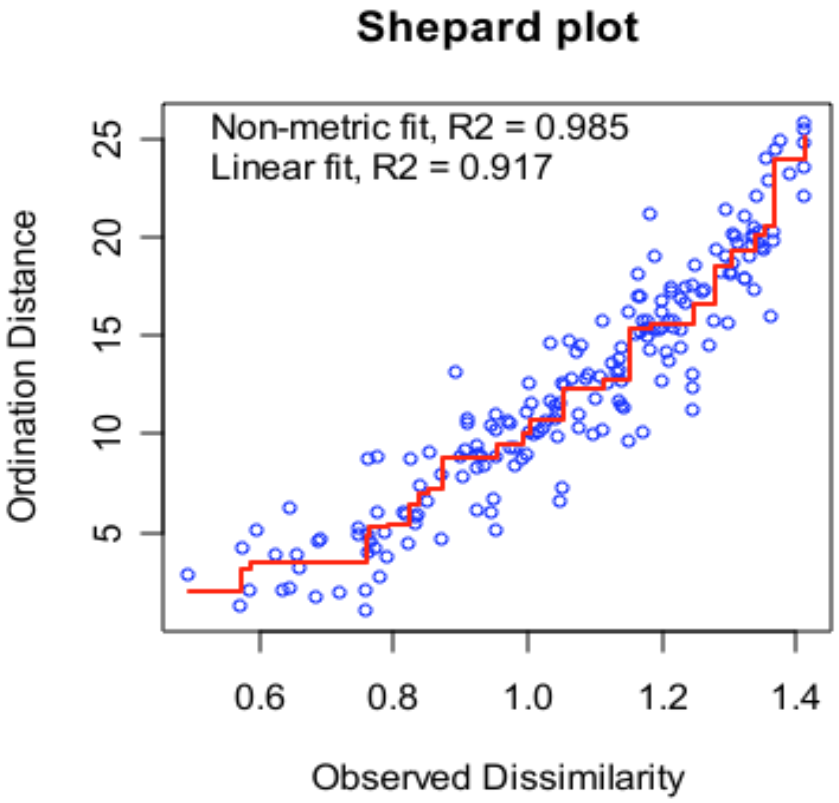
- Is based on any dissimilarity or distance matrix (even non-metric) and tolerates missing data
- Non-parametric method which preserves the order of resemblance of the objects (ranks) along a few (usually 2) axes
  - Result depends on the number of axes selected
- The position of the origin, scale and orientation of the axes are arbitrary
  - distances between objects is not a priority
  - axes are not hierarchical
  - axes may undergo inversion, rotation or re-centering

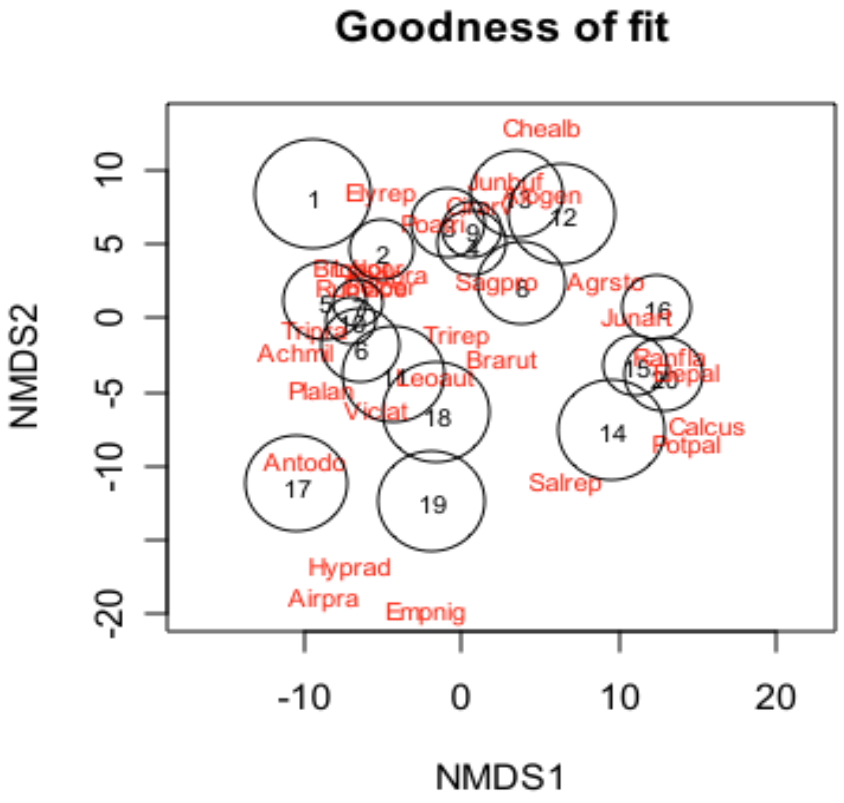
Stages of NMDS

homework

1. Select a the distance matrix (e.g. Bray-Curtis)
2. Select the number of axes (default: 2)
3. Selection of the initial configuration of the objects (random or PCoA)
4. Calculation of the matrix of distances d_{hi} between objects of this initial configuration
5. Representation of the Shepard diagram and regression of the adjusted distances with the observed distances
6. Measure of the *goodness-of-fit* of the regression with an objective function of *stress*
7. Improvement of the configuration in order to minimize the stress function
8. Iteration of the stages 4 to 7 until convergence, which then produces the coordinates of the objects in the ordination

Shepard diagram and quality of the adjustment





Guideline:

stress < 0.05 excellent

stress < 0.1 good

stress < 0.2 usable

stress > 0.2 not acceptable

EPFL

Application of NMDS to species abundance data…

- Very robust and flexible non-parametric method, advocated by some authors
- NMDS is an iterative optimization method and results depend on the number of axes and the initial configuration
  - It is advisable to test different initial configurations
- Recentering and rotation of the axes is done in order to maximize the variance on the first axis
- Possibility to project the species *a posteriori* on the NMDS plot

Choice of (unconstrained) ordination

1. PCA: quantitative descriptors (or semi-quantitative)
    1. Covariance or correlation according to data (similar scale) transformation: in case of long gradients or many zeros

2. CA: descriptors of any kind but of *same nature*
    1. Remove rare species or/and transformation of data in case of horseshoe effects

3. PCoA and NMDS: choice of association metric
    1. PCoA requires a metric and Euclidean distances matrix
    2. NMDS is more flexible but requires precautions to ensure convergence to an optimal solution

Some general advice on ordination

1. Objects (samples) should be independent
2. If possible, have more objects than descriptors
3. Consider at least axes 1-2 and 1-3 and take into account the percentage of variance represented on each axis (screeplots)
4. Superimpose the results of an unsupervised cluster analysis or *a priori* known groups
5. Interpret the axes by an indirect gradient analysis (e.g. correlate scores with an env. parameter)