„Applied“ Multivariate Statistics with R

for ICBM R Roundtable

March 2017

Helena Osterholz

1. Why multivariate statistics?

- to avoid multiple testing!
- to show variation in data on a reduced number of dimensions

The aim of ordination is the representation of similarity between objects (samples, sites) based on values of multiple variables (species, measured parameters) asscociated with them.

Before you start...

...familiarize yourself with your data!

- use histograms, scatterplots, boxplots, etc...
- missing datapoints?
- outliers?
- identifiy highly correlated variables
- linear/unimodal distributions?

Example: corrplot

Prepare data

Standardization

- ecological data usually don‘t vary on the same scales (e.g. salinity, cell counts) and should be scaled for better comparison

  →use normalizing transformations, like Z-scoring → assumes that the mean is a good representation of the data)

Normalization

- corrects the distribution of variables that depart from normality
- to make species data containing many 0s suitable for linear analysis: Hellinger transformation

Example: decostand, scale, shapiro.test

2. Data Exploration

Exploratory analyses reveal patterns in datasets, but do not explain why those patterns exist.

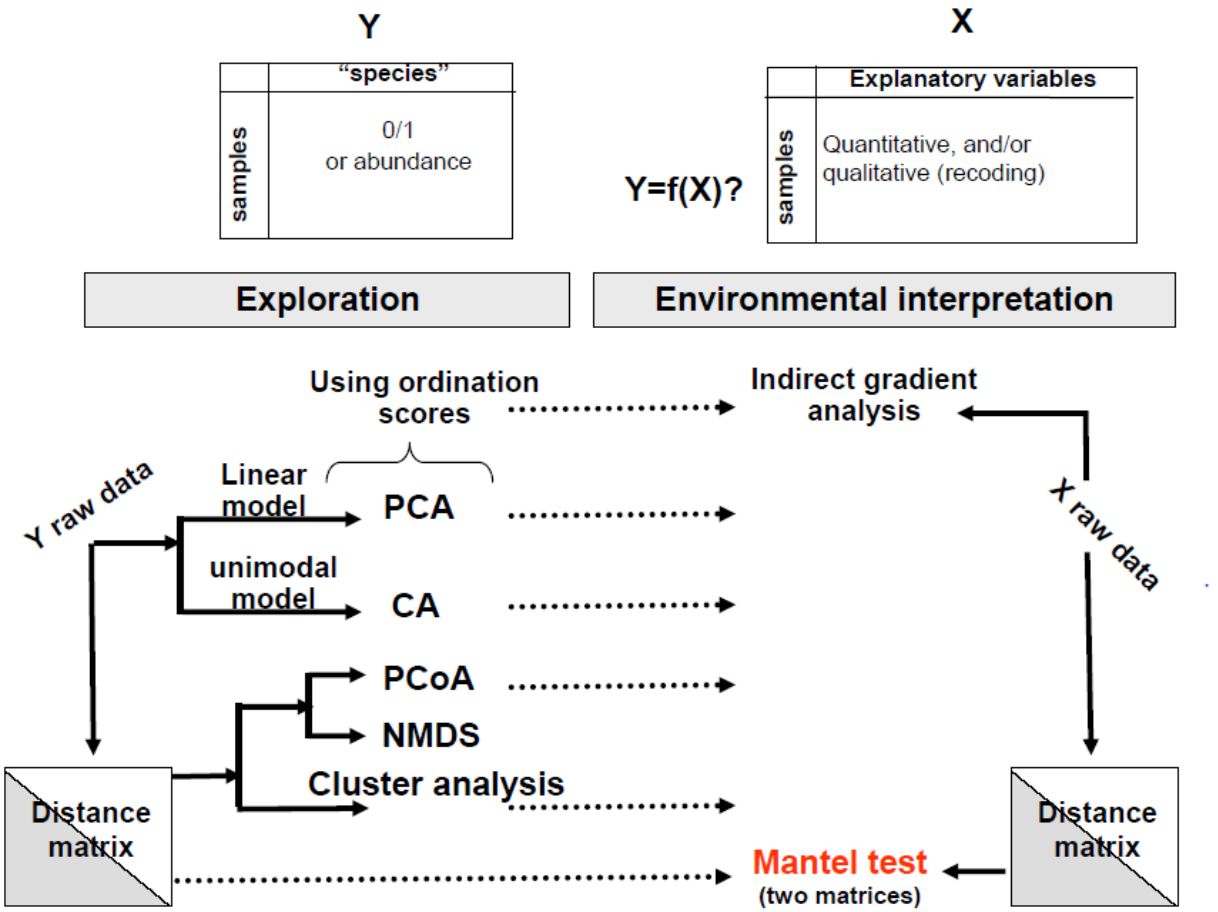
...if relationships are discontinuous:

- Cluster analysis

...if relationships are related to a gradients:

- NMDS (non-metric)
- PCA, PCoA (metric)

Choosing ordinations



from A. Ramette

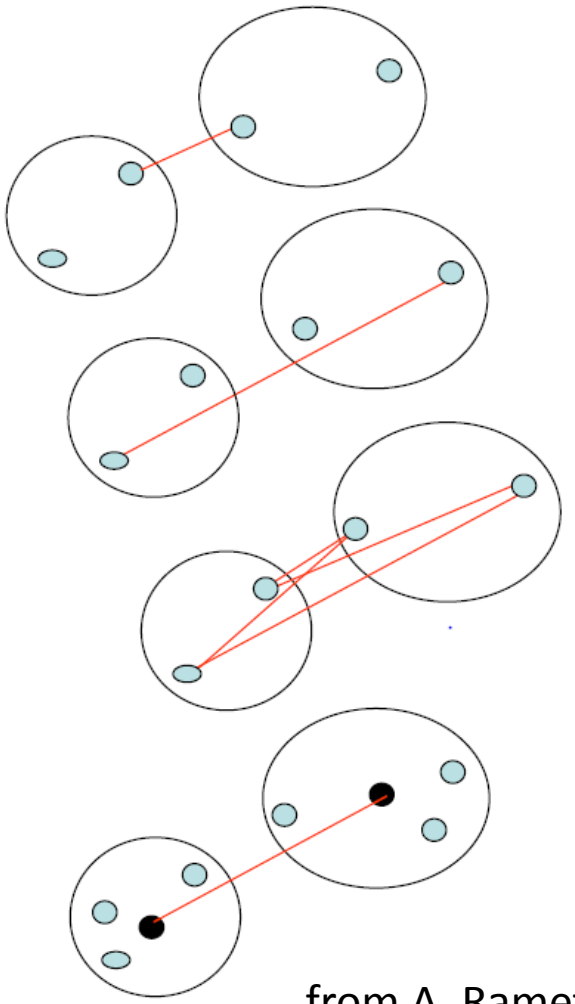
Cluster Analysis

- Finds hierarchical groupings in multivariate datasets
- recommended when distinct disontinuities instead of gradients are expected
- 2 steps:
  - 1. calculate distance matrix
  - 2. represent tree (hierarchical clustering) or clusters (k-means clustering)

Example: hclust

Cluster Analysis: Methods

- **nearest neighbour**: the distance between two clusters is the distance between their closest neighbouring points
- **furthest neighbour**: the distance between two clusters is the distance between their two furthest objects
- **UPGMA**: unweighted pair-group method using averages
- **centroid method**: uses the centroid to determine the average distance between clusters
- **Ward's method**: when within-cluster homogeneity is desired.
- Neighbour joining cluster analysis – in contrast to UPGMA, two banches from the same internal node do not need to have equal branch length



from A. Ramette

Cluster Analysis

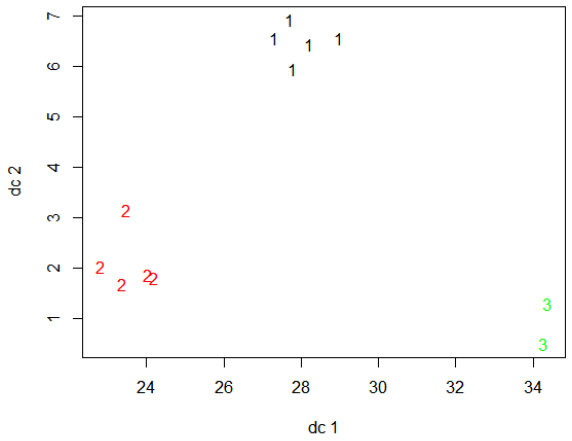
- **Bootstrapping:**

random resampling of columns with replacement, each time calculate a cluster analysis, for each cluster – determine how many dendrograms out otf the n repetitions lead to the formation of that specific cluster

Example: clusterboot

K-means clustering

- a priori definition of the number of clusters
- very sensitive towards outliers
- usual k-means algorithm uses Euclidean distance



Discriminant projection plot

Example: kmeans

Non-metric Multidimensional Scaling [NMDS]

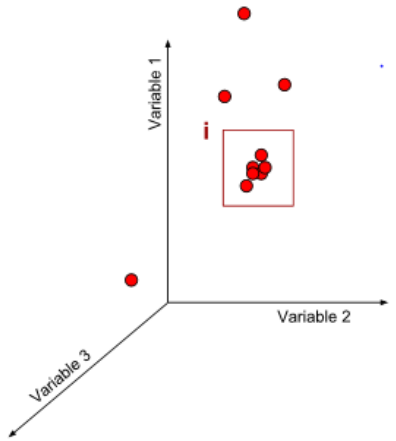
- distances between objects are ranked
- ranks are used to map the objects (samples) non-linearly onto a 2-dimensional ordination space, but do not correpond to original distances between objects
- axes can be freely rescales, rotated or inverted

→ efficient at identifying underlying gradients and representing relationships based on different distance measures

Example: metaMDS, varimax

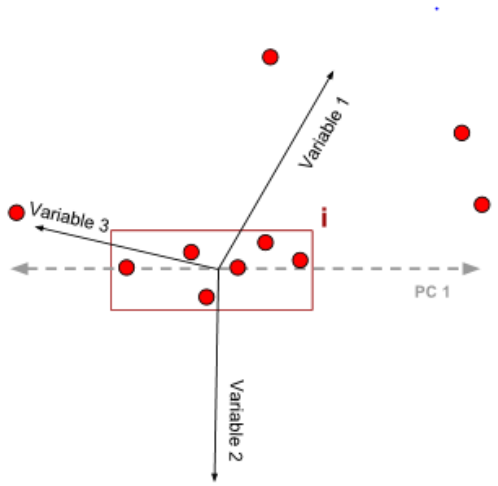
Principal Component Analysis [PCA]

- calculates new synthetic variables (principal components) which are linear combinations of the original variables and which account for as much of the variance of the original data as possible
- rigid rotation of the original system of axes: the successive new axes are orthogonal to one another



Disadvantages of PCA:

- only works for quantitative data
- not too many zeros should be present
- assumption: linear relationship of each variable with the environment or its components



Example: prcomp, evplot, scree

from Gusta Me

Principal Coordinate Analysis [PCoA]

- PCoA
- uses linear (Euclidean) mapping of the distance or dissimilarities between objects
- works with any dissimilarity measure
- no direct link between the components and the original variables, so interpretation of variable contribution may be more difficult (no loadings)

Addition of environmetal parameters:

- fitting vectors: arrow point tot he direction of the most rapid change in the environmental variable → direction of the gradient
- length of the arrow is proportional to the correlation between ordination and environmental variable → strength of gradient
- plotting of isolines (recommended by some over vector fitting)

Example: cmdscale, envfit, ordisurf

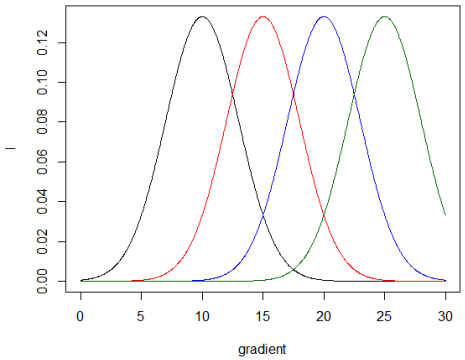
Correspondence Analysis [CA]

- to compare correspondence between samples and species from a e.g. table of counted data
- assumes unimodal distribution of species (one optimal condition exists along the gradient for a given species)
- In CA biplot, proximity can be understood as a probablility of occurence or high abundance of a given species in a sample.
- CA is sensitive to rare species.
- CA tends to form arches or horseshoe-like plots.

Detrending can be perfomed to create a linear mapping in such a case (DCA)

– not recommended.

Example: ca



PCA vs CA

- main difference: how the variation in species is quantified before weights are assigned

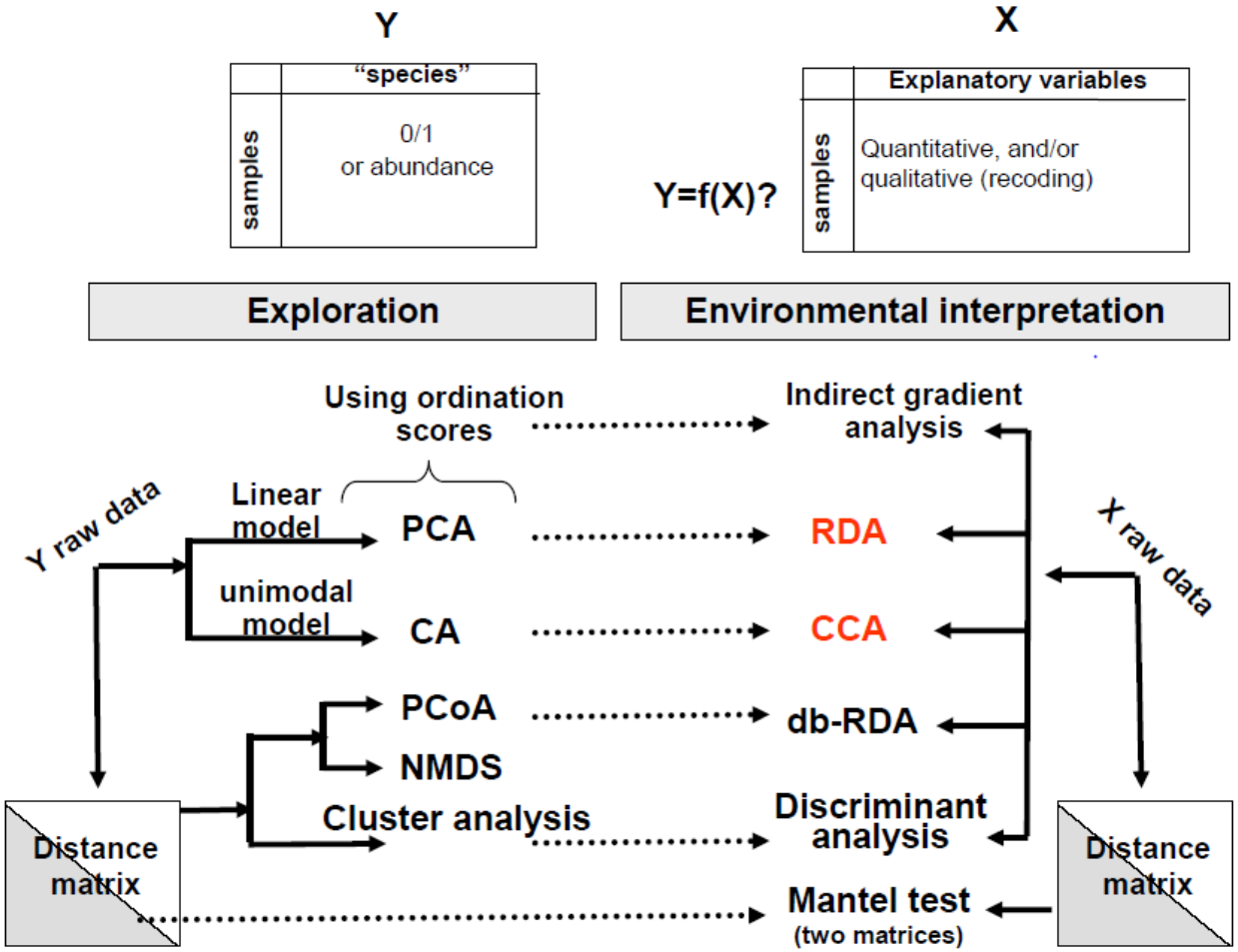
- PCA: linear model. Species abundances → cov or corr matrix → eigenanalysis to determine weighted combinations
- CA: unimodal model. Species abundances → chi-square distances → variance is evaluated by eigenanalysis.

Presumably, CA preserves ecological distance by modeling differences in associations rather than abundances of single species.

3. Constrained Analyses

Aim: to find mathematical relationships between species composition and a measured environmental variable, to assess statistical significance of the relationship, and to represent those relationships in low-dimensional space.

- CanCor
- CCA
- RDA
- variation partitioning



from A. Ramette

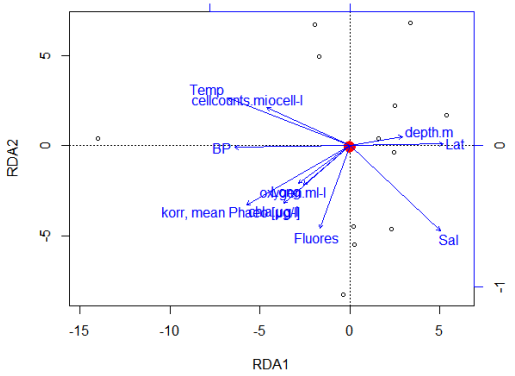
Redundancy Analysis [RDA]

- basically a PCA, but axes are restricted to be linear combinations of explanatory variables
- models **linear** species-environmental relationships

RDA can be represented by a triplot with samples (dots), species (arrows) and environmetal variables (arrows for quantitative variables, dots for qualitative variables).

Scaling 1: focus on intersample relationships

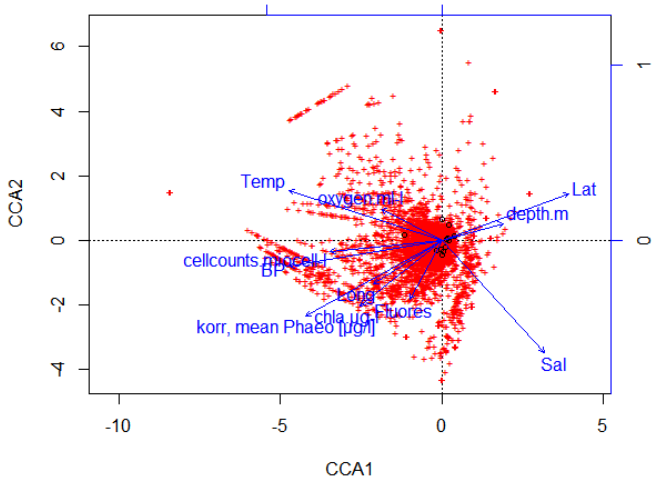
Scaling 2: focus on interspecies correlations



Example: rda

Canonical Correspondance Analysis [CCA]

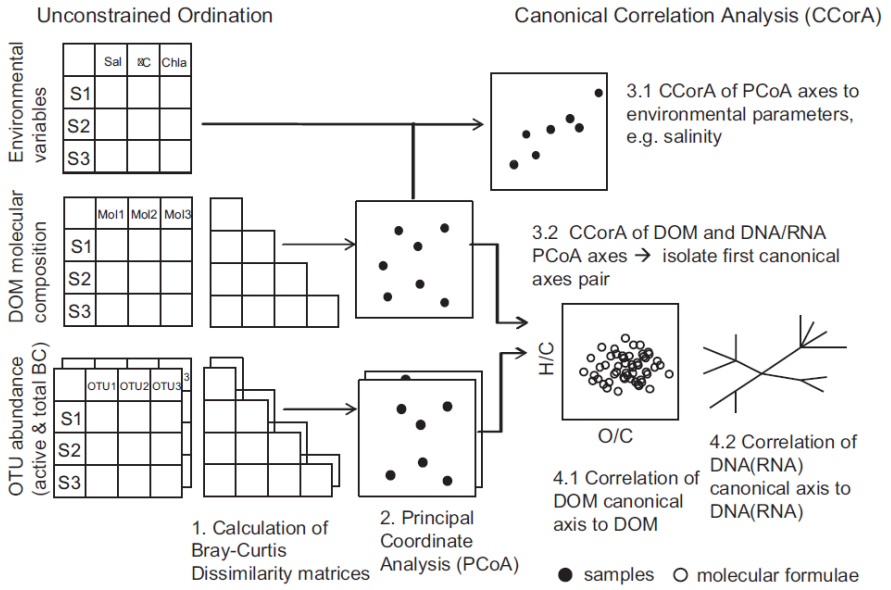
- based on unimodal species-environment relationships
- canonical form of CA
- works well with many 0s
- sensitive to rare species



Example: cca

Canonical Correlation Analysis [CCorA]

- bimultivariate approach that can be used in case it is unclear which variables are explanatory/response (symmetrical analysis)
- linear method



Example: CCorA

Mantel Test & Procrustes Rotation

Mantel

- test for global association of two matrices:
- „Do pairs of sites with similar environmetal variables also harbor a similar species composition?“
- calculates the correlation coefficient r between corresponding positions in the two matrices
- has a partial option

Procrustes

- compares multidimensional shapes by attempting to minimise the sum of squared differences

Example: mantel, partial.mantel, protest

Reducing the number of environmental variables

- if response and explanatory variables are linearly related, exhibit homoscedasticity, residuals are normally distributed,...
- very susceptible to outliers
- selection can be unstable if explanatory variables are multicollinear

- variance inflation facors VIF to test for multicollinearity of explanatory variables – rule of thumb: should be <10

Example: ordistep, vif.cca

Variation partitioning

- E.g. an environmental gradient is known to occur, partial ordination can be used to investigate the effects of other variables or cominations, while taking into consideration this gradient.

dbRDA

- taking into account „space“ as an environmental variable

Useful resources

- https://mb3is.megx.net/gustame
- http://ordination.okstate.edu/
- Legendre & Legendre (2012), Numerical Ecology, 3rd edition, Elsevier B.V., Amsterdam, Netherlands

...correlation does not imply causation...