

# Tools for complex data analysis

University of Koblenz-Landau 2018/19



Ralf B. Schäfer

# **RDA, similarity measures and NMDS**

## **Contents**

- 1. Learning targets, constrained ordination  
and RDA**
2. Diagnosis and assumptions of RDA,  
extensions and guidance for method  
selection
3. Similarity and distance measures
4. Non-metric multidimensional scaling (NMDS)

# Learning targets

- Understanding the basics of RDA.
- Knowledge on the calculation of commonly used association measures.
- Understanding their suitability for ecological data.
- Understanding the mathematical background and how to conduct a NMDS.

# Learning targets and study questions

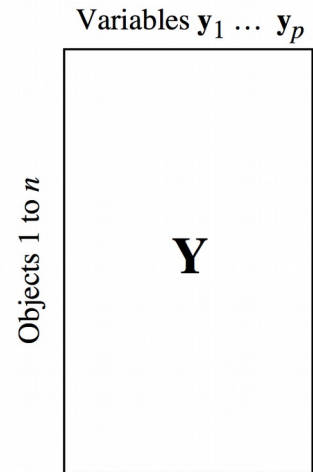
- Understanding the basics of RDA.
  - How many constrained axes has an RDA and how are they related to the descriptors?
  - How does scaling influence the interpretation of a triplot?
- Knowledge on the calculation of commonly used association measures.
  - Which association is measured with similarity measures?
  - Outline the calculation of the Bray-Curtis and the Jaccard coefficient.
- Understanding their suitability for ecological data.
  - Explain the double-zero problem.
  - What is the species abundance paradox?

# Learning targets and study questions

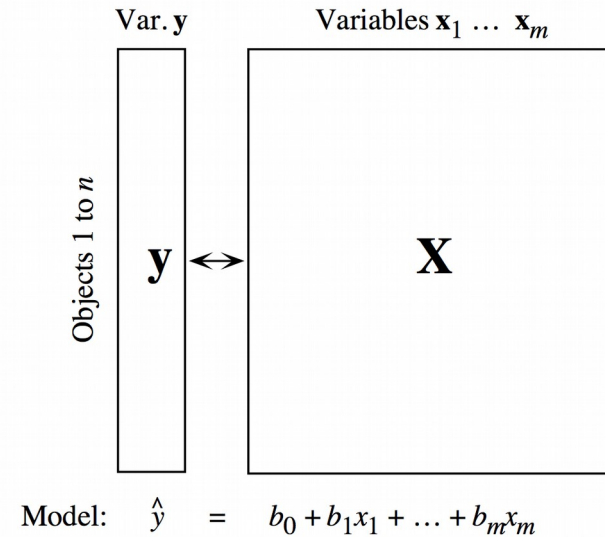
- Understanding the mathematical background and how to conduct a NMDS.
  - What are the main differences between NMDS and PCA?
  - Which three matrices are computed during NMDS?
  - Outline the major elements of the algorithm used to compute the NMDS.
  - Discuss limitations of NMDS.

# Constrained ordination methods

(a) Simple ordination of matrix **Y**:  
principal comp. analysis (PCA)  
correspondence analysis (CA)

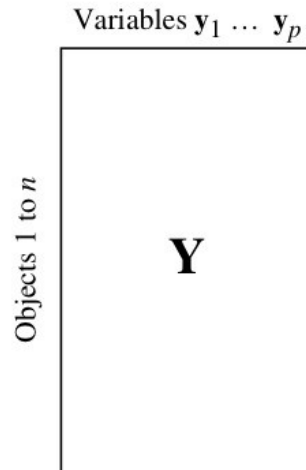


(b) Ordination of **y** (single axis) under  
constraint of **X**: multiple regression

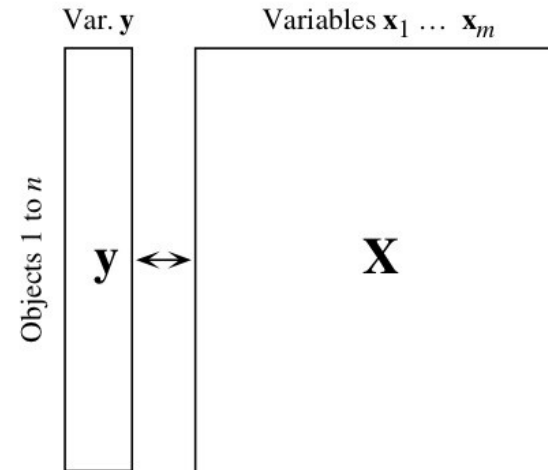


# Constrained ordination methods

(a) Simple ordination of matrix **Y**:  
principal comp. analysis (PCA)  
correspondence analysis (CA)

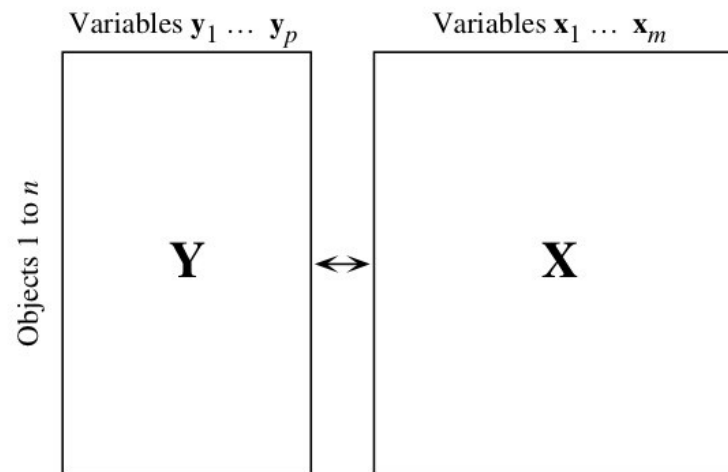


(b) Ordination of **y** (single axis) under  
constraint of **X**: multiple regression

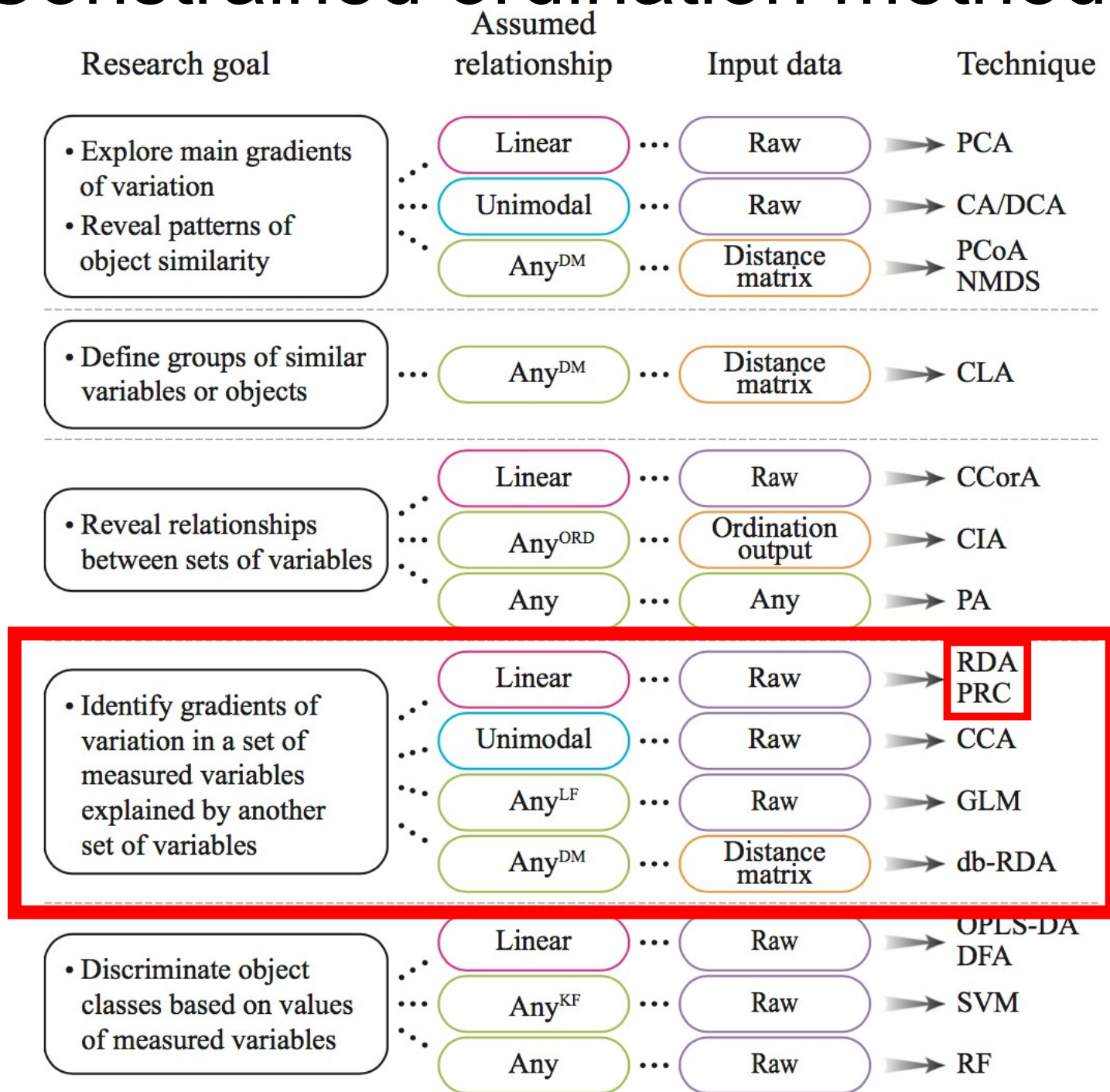


$$\text{Model: } \hat{y} = b_0 + b_1x_1 + \dots + b_mx_m$$

(c) Ordination of **Y** under constraint of **X**:  
redundancy analysis (RDA)  
canonical correspondence analysis (CCA)



# Constrained ordination methods

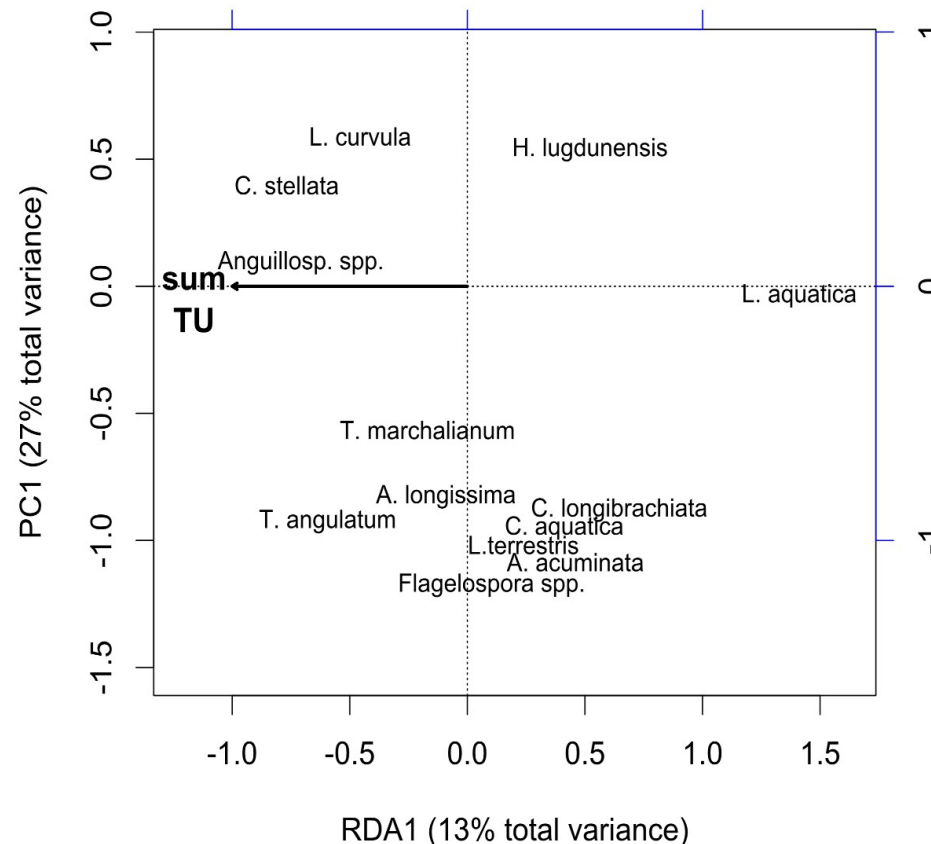




# Redundancy Analysis (RDA)

**Aim:** Display and explain variation in set of response variables constrained by second set of predictor variables  
→ Links multivariate multiple regression and PCA

**Example:** Which variable(s) do best explain the variation in fungal communities sampled along a gradient of fungicide toxicity?




# Mathematical background of RDA

**Aim:** Display and explain variation in set of response variables constrained by second set of predictor variables  
→ Links multivariate multiple regression and PCA

Remember: Multiple linear regression in matrix form

$$\begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{pmatrix} \quad \Rightarrow \quad \hat{y} = \mathbf{X} b$$



$$b = (\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X}^T y)$$

Substitution yields:  $\hat{y} = \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X}^T y)$

Reformulation for multivariate multiple regression with several  $y$ :

$$\hat{\mathbf{Y}} = \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{Y})$$


# Mathematical background of RDA

$$\hat{Y} = X(X^T X)^{-1}(X^T Y)$$

RDA uses variance-covariance matrix of  $\hat{Y} \Rightarrow \Sigma_{Y^T Y}$

Usually, this is not known and the sample variance-covariance matrix (also called Dispersion matrix) is estimated from the observations:

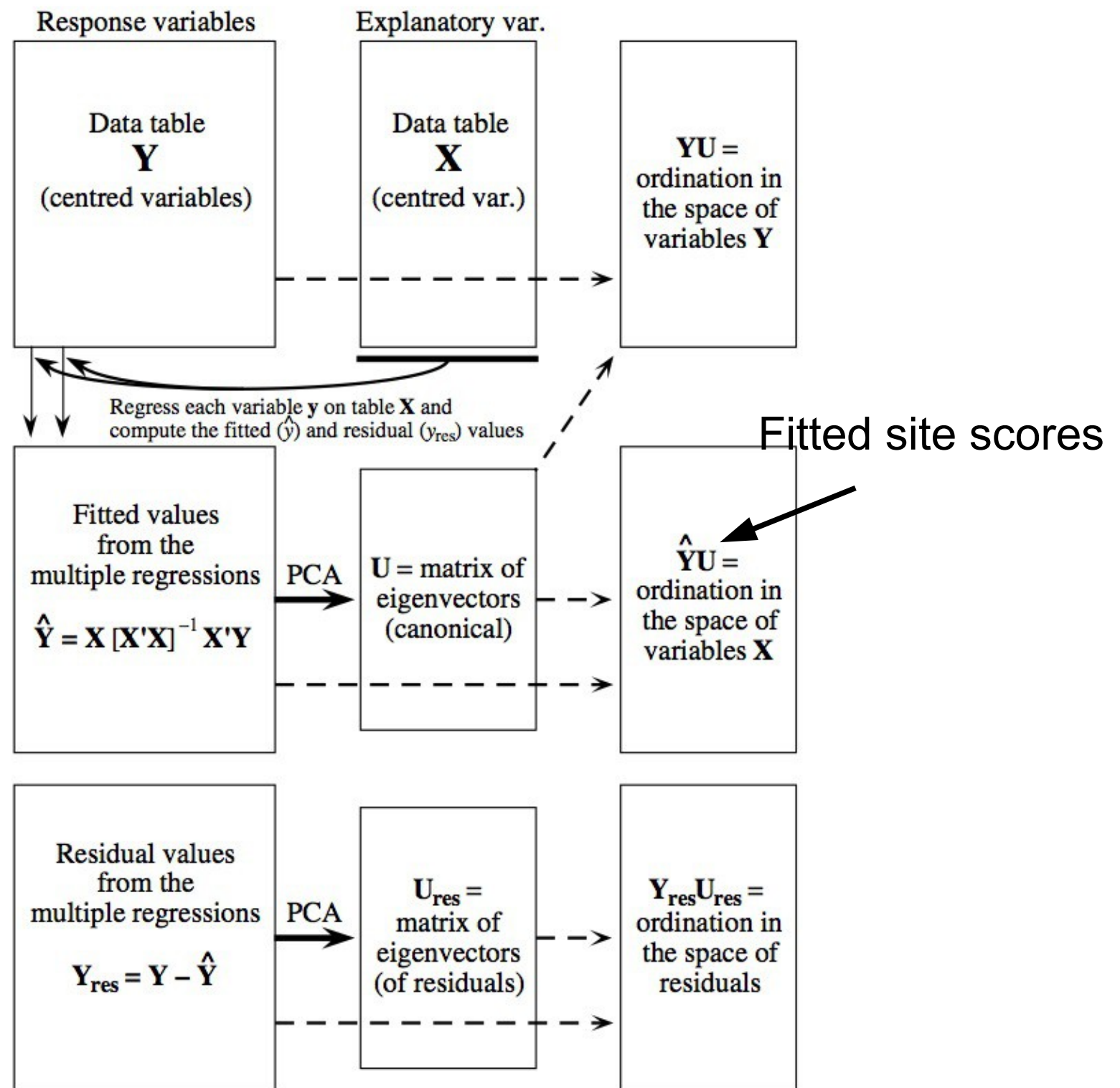
$$S_{\hat{Y}^T \hat{Y}} = \frac{1}{n-1} \hat{Y}^T \hat{Y}$$

and used in a PCA:  $S_{\hat{Y}^T \hat{Y}} a = \lambda a$   Eigenvector

Eigenvalue problem

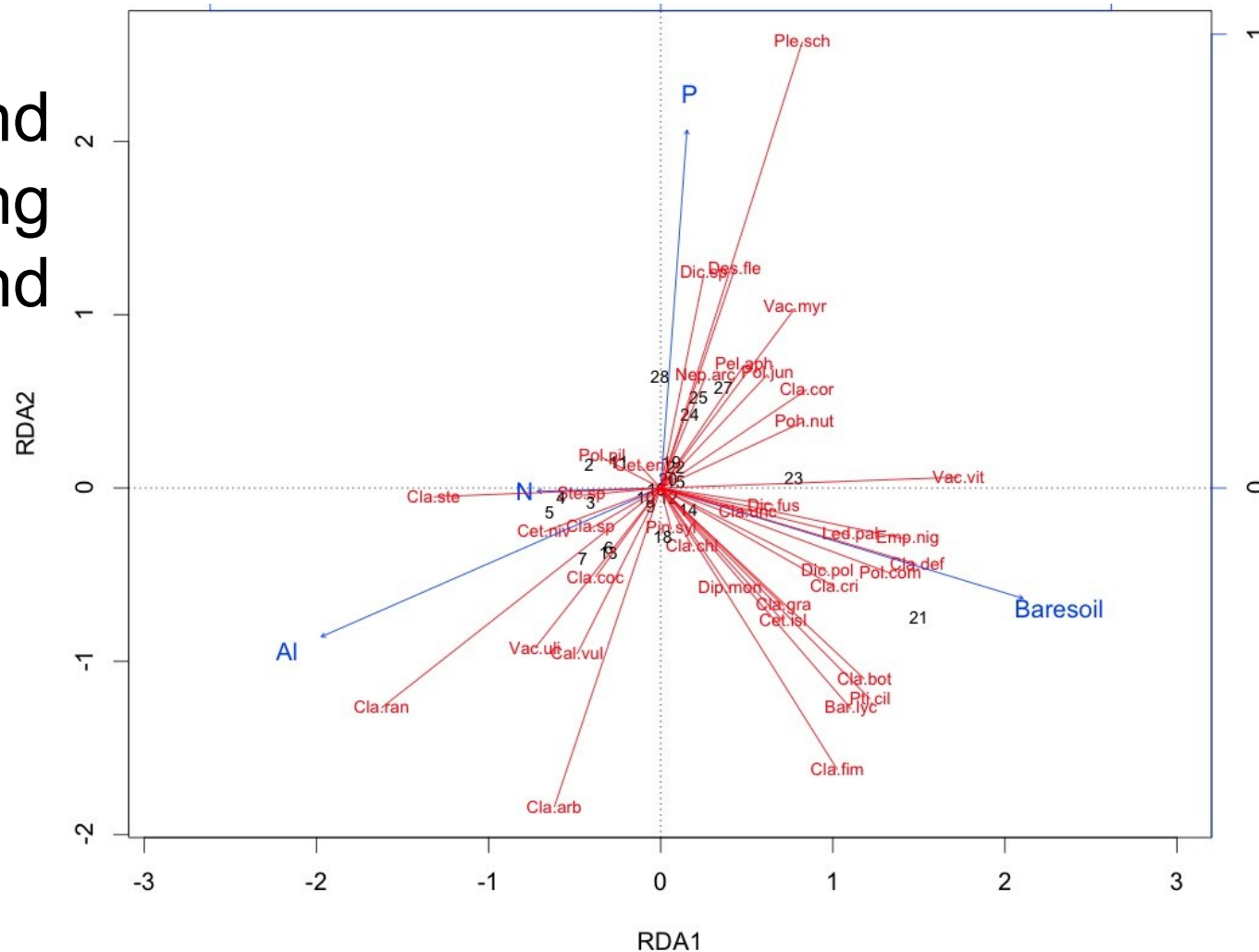


Eigenvectors linear combinations of predictors



# RDA results

- Triplot with relationship between species, sites and env. variables
- Eigenvalues and variance partitioning (constrained and unconstrained)
- Site scores
- Species scores
- Biplot scores for variables



# RDA, similarity measures and NMDS

## Contents

1. Learning targets, constrained ordination and RDA
- 2. Diagnosis and assumptions of RDA, extensions and guidance for method selection**
3. Similarity and distance measures
4. Non-metric multidimensional scaling (NMDS)

# RDA axes and variable importance

How many RDA axes are required?

- Hypothesis test (permutation-based) recommended (Legendre et al. *Methods Ecol. Evol.* 2011)

How many environmental variables are needed and how important are they?

- Manual and automatic model-building with *adj.  $R^2$*  as goodness of fit criteria (as for multiple linear regression)
- Variance partitioning between different models to determine explained variance of individual variables

# Assumptions and extensions of RDA

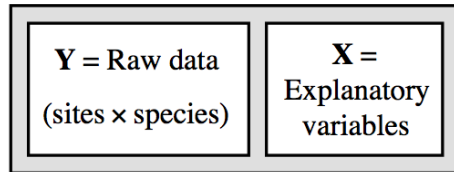
- Independence of observations (sites)
- Linear relationship between explanatory and response variables → see next slide
- No multicollinearity between explanatory variables
- $n$  (sites)  $\gg p$  (predictors) to reliably infer  $p$  importance
- RDA can be employed for multivariate ANOVA (see Borcard et al. 2011: 185 ff)
- RDA over time important for ecotoxicological experiments:  
→ Principal Response Curves (PRC) that deliver time-dependent treatment effects relative to control (van den Brink & ter Braak 1999 *ET&C* 18 (2): 138-148)



# RDA approaches

## How to assess gradient length?

(a) Classical approach: RDA preserves the Euclidean distance, CCA preserves the chi-square distance



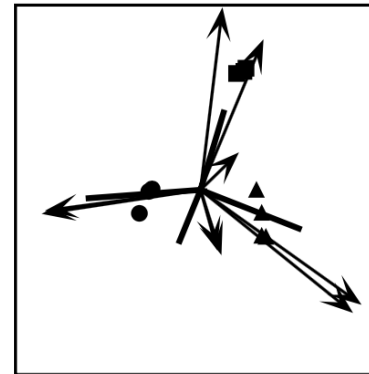
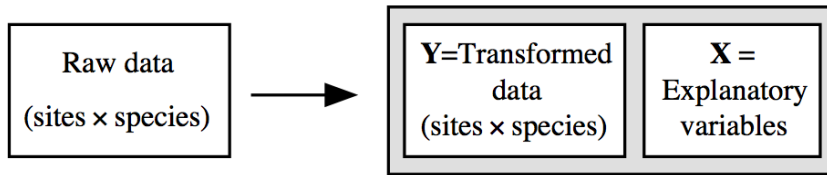
Short gradients: CCA or RDA

Long gradients: CCA

- test for higher order terms (Borcard et al. 2011: 190ff)
- Axis length in DCA

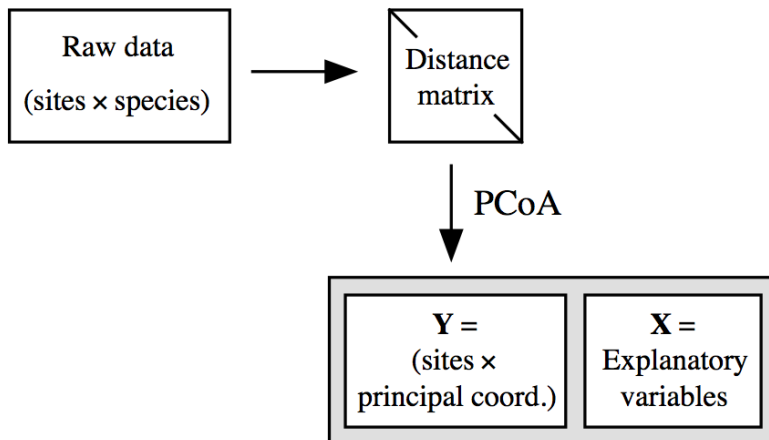
Canonical  
ordination triplot

(b) Transformation-based RDA (tb-RDA) approach:  
preserves a distance obtained by data transformation



Representation of elements:  
Species = arrows  
Sites = symbols  
Explanatory variables = lines

(c) Distance-based RDA (db-RDA) approach:  
preserves a pre-computed distance



# Further constrained ordination methods

## Canonical Correspondence Analysis (CCA)

- Widely used
- Extension of (unconstrained) correspondence analysis
- Similar to RDA, but assumes unimodal distribution ( $\chi^2$ -distance) of species along environmental gradient
- In R: model building as for RDA

`cca()` {vegan}

## Constrained additive Ordination (CAO)

- Comparatively new
- derives response of each species to main environmental gradient from data → no linear or unimodal model assumed
- mixture of Generalized Additive Models (GAMs) and Canonical Gaussian Ordination
- computationally demanding
- In R: implemented in extra package

`cao()` {VGAM}

# When to use what?

---

Numerical methods to *forecast* one or several descriptors (response or dependent variables) using other descriptors (explanatory or independent variables). In parentheses, identification of the section where a method is discussed.

---

- 1) Forecasting the structure of a *single* descriptor, or *indirect comparison* ..... see 2
  - 2) The response variable is quantitative ..... see 3
    - 3) The explanatory variables are quantitative ..... see 4
      - 4) Null or low correlations among explanatory variables: *multiple linear regression* (10.3); *nonlinear regression* (10.3)
      - 4) High correlations among explanatory variables (collinearity): *ridge regression* (10.3); *regression on principal components* (10.3)
    - 3) The explanatory variables are qualitative: *dummy variable regression* (10.3)
  - 2) The response variable is qualitative (*or* a classification) ..... see 5
    - 5) Response: two or more groups; explanatory variables are quantitative (but qualitative variables may be recoded into dummy variables): *identification functions in discriminant analysis* (11.3)
    - 5) Response: binary (presence-absence); explanatory variables are quantitative (but qualitative variables may be recoded into dummy var.): *logistic regression* (10.3)
  - 2) The response and explanatory variables are quantitative, but they display a nonlinear relationship: *nonlinear regression* (10.3)
- 1) Forecasting the structure of a *multivariate* data matrix ..... see 6
  - 6) *Direct comparison* ..... see 7
    - 7) Linear modelling: *redundancy analysis* (RDA, 11.1); *canonical correspondence analysis* (CCA, 11.2)
    - 7) Find a tree-like decision model: *multivariate regression tree analysis* (MRT, 8.11)
  - 6) *Indirect comparison* ..... see 8
    - 8) Ordination in reduced space: each axis is treated in the same way as a single quantitative descriptor ..... see 2
    - 8) Clustering: each partition is treated as a qualitative descriptor ..... see 2





# RDA, similarity measures and NMDS

## Contents

1. Learning targets, constrained ordination and RDA
2. Diagnosis and assumptions of RDA, extensions and guidance for method selection
- 3. Similarity and distance measures**
4. Non-metric multidimensional scaling (NMDS)

# Measuring association

## Example: Species observations in 4 streams

Site				
1	0	400	0	0
2	0	0	10	0
3	2	280	3	3
4	12	60	80	50

**What is the relationship between 1) objects 2) descriptors?**

- Relationship between objects (sites): distance or similarity measures
- Relationship between descriptors (species): Dependence measures (e.g. covariance or correlation between environmental variables)





# Similarity measures: Presence-absence

## Simple matching coefficient





		Site 1		
		present	absent	
Site 2	present	a	b	a + b
	absent	c	d	c + d
Sum		a + c	b + d	

$$S_m = \frac{a + d}{a + b + c + d}$$

Exercise: Calculate  $S_m$  for the data below with and without the 1. and 4. species. How do these species influence  $S_m$ ?

Site				
1	0	400	0	0
2	0	0	10	0

# Similarity measures: Presence-absence

Site				
1	0	400	0	0
2	0	0	10	0

$$S_m = \frac{a + d}{a + b + c + d}$$

Calculation with all species:

$$a = 0, b = 1, c = 1, d = 2 \rightarrow S_m = 2/4 = 0.5$$

Calculation without species 1 and 4:

$$a = 0, b = 1, c = 1, d = 0 \rightarrow S_m = 0/2 = 0$$

Species absence influences similarity between sites.

Not desirable: joint absence of species does not indicate ecological similarity and number of joint absences is arbitrary

→ **Double-Zero problem**

# Widely used similarity measures

## Jaccard coefficient (=Jaccard similarity index)

		Site 1		
		present	absent	
Site 2	present	a	b	a + b
	absent	c	d	c + d
Sum		a + c	b + d	

$$S_j = \frac{a}{a+b+c}$$

- used for binary data
- ignores joint absences (d)





## Bray-Curtis coefficient

- used for abundance data
- range: 0 - 1 (if all  $x_k \geq 0$ )
- data transformation often required to reduce weight of dominant taxa

$$S_{BC}(i, j) = \frac{2 \sum_{k=1}^n \min(x_{i,k}, x_{j,k})}{\sum_{k=1}^n |x_{i,k} + x_{j,k}|}$$



# Example: Bray-Curtis coefficient

Site				
1	0	400	5	0
2	0	0	10	0
Min	0	0	5	0
Sum	0	400	15	0

$$S_{BC}(i, j) = \frac{2 \sum_{k=1}^n \min(x_{i,k}, x_{j,k})}{\sum_{i=1}^n |x_{i,k} + x_{j,k}|}$$

Calculation:

$$2 \cdot (0 + 0 + 5 + 0) / 415 \rightarrow S_{BC} = 10 / 415 = 0.025$$

Calculation after square-root transformation:

$$2 \cdot (0 + 0 + 5^{0.5} + 0) / (400^{0.5} + 5^{0.5} + 10^{0.5}) \rightarrow S_{BC} = 0.18$$

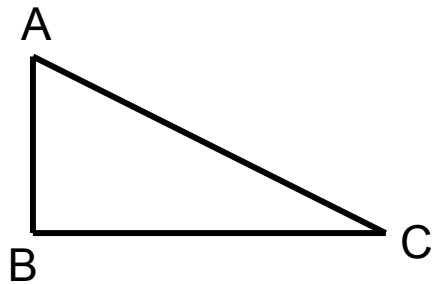
Calculation after double square-root transformation:

$$2 \cdot (0 + 0 + 5^{0.25} + 0) / (400^{0.25} + 5^{0.25} + 10^{0.25}) \rightarrow S_{BC} = 0.39$$

# Distance measures

## Association measures meeting triangle inequality criterion

(following Everitt et al. 2011 *Cluster Analysis*. John Wiley & Sons: 49)



### Triangle inequality criterion

$d(A,B) + d(B,C) \geq d(A,C)$ , where  $d$  is distance function

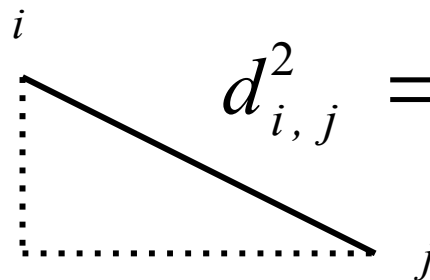
Sum of any two sides of triangle always  $\geq$  third side

Important for geometrical representation (e.g. ordination)

Euclidean distance: Most frequently used distance measure

$$d_{i,j} = \sqrt{\sum_{k=1}^n (x_{i,k} - x_{j,k})^2}$$

Two dimensional case:



$$d_{i,j}^2 = (x_{i,1} - x_{j,1})^2 + (x_{i,2} - x_{j,2})^2$$

# Species abundance paradox

**Species x Site matrix**

Sites	Species		
	$y_1$	$y_2$	$y_3$
$x_1$	0	1	1
$x_2$	1	0	0
$x_3$	0	4	4

Euclidean  
distance



**Distance matrix**

Sites	Sites		
	$x_1$	$x_2$	$x_3$
$x_1$	0	1.732	4.243
$x_2$	1.732	0	5.745
$x_3$	4.243	5.745	0

Sites  $x_1$  and  $x_2$  share no species, but have a smaller distance than sites sharing species ( $x_1$  and  $x_3$ ).

→ Euclidean distance problematic for ecological data

# How to select a measure

- Many more association measures  
(see Legendre & Legendre 2012: Chapter 7)
- Check literature of scientific field
- Refer to key in Legendre & Legendre 2012: 325-328

Choice of an association measure among objects (Q mode), to be used with chemical, geological physical, etc. descriptors (symmetrical coefficients, using double-zeros).

---

- |  |       |
|--|-------|
| 1) Association measured between individual objects   | see 2 |
| 2) Descriptors: presence-absence or multistate (no partial similarities computed between states)                 | see 3 |
| 3) Metric coefficients: <i>simple matching</i> ( $S_1$ ) and derived coefficients ( $S_2, S_6$ )                 |       |
| 3) Semimetric coefficients: $S_3, S_5$   |       |
| 3) Nonmetric coefficient: $S_4$  |       |
| 2) Descriptors: multistate (states defined in such a way that partial similarities can be computed between them) | see 4 |
| 4) Descriptors: quantitative and dimensionally homogeneous   | see 5 |
| 5) Differences enhanced by squaring: <i>Euclidean distance</i> ( $D_1$ ) and <i>average distance</i> ( $D_2$ )   |       |

# Association measures in R

Function	Package	No. of measures	Weighing possible?
dist	stats	6	No
daisy	cluster	3	Yes
dsvdis	labdsv	7	Yes
vegdist	vegan	14 (easily expandable)	No
distance	ecodist	10 (easily expandable)	Yes
dist.*	ade4	~25 (in different functions)	Yes

# RDA, similarity measures and NMDS

## Contents

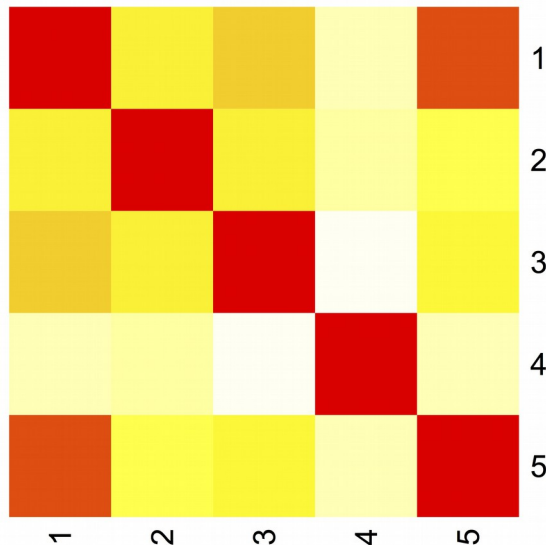
1. Learning targets, constrained ordination and RDA
2. Diagnosis and assumptions of RDA, extensions and guidance for method selection
3. Similarity and distance measures
- 4. Non-metric multidimensional scaling (NMDS)**

# Visualization of association measures

## Heatmap

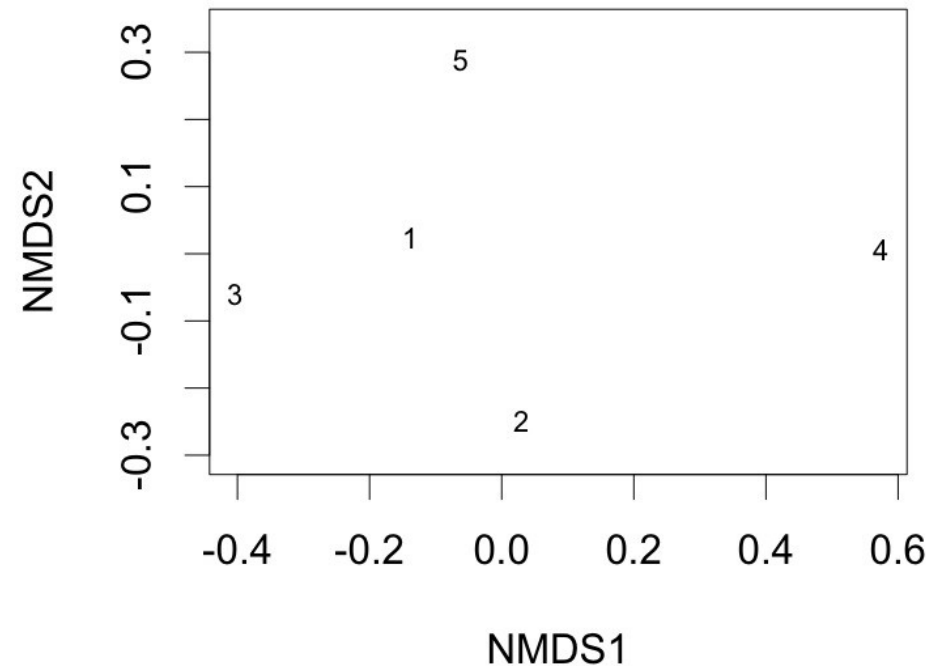
- Associations converted to colours
- Relationship easier to grasp

	1	2	3	4	5
1	0.00	0.69	0.60	0.92	0.22
2	0.69	0.00	0.70	0.89	0.80
3	0.60	0.70	0.00	0.98	0.72
4	0.92	0.89	0.98	0.00	0.92
5	0.22	0.80	0.72	0.92	0.00

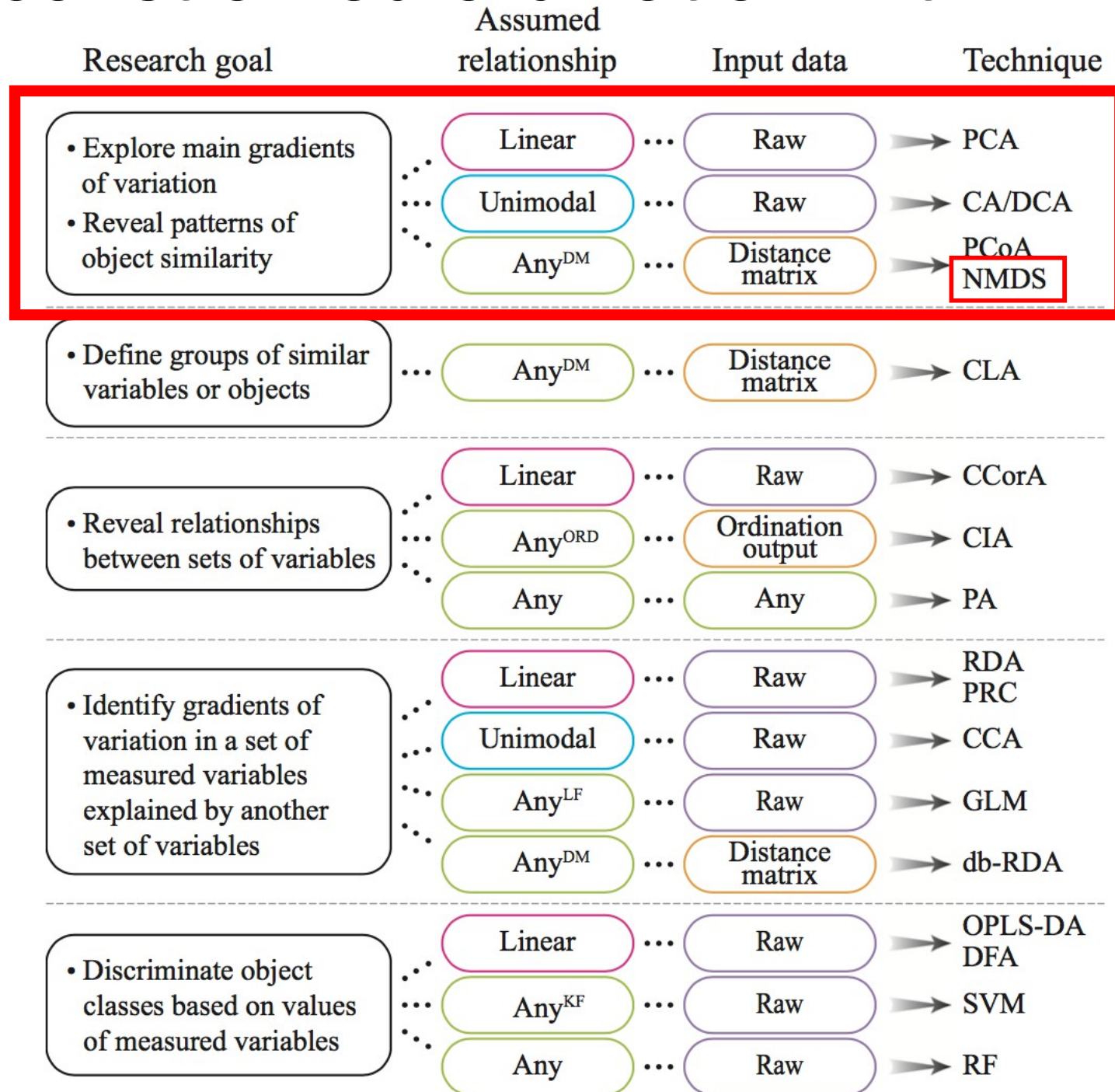


## Ordination

- Works for measures that meet triangle inequality criterion (otherwise no clear geometrical interpretation possible)



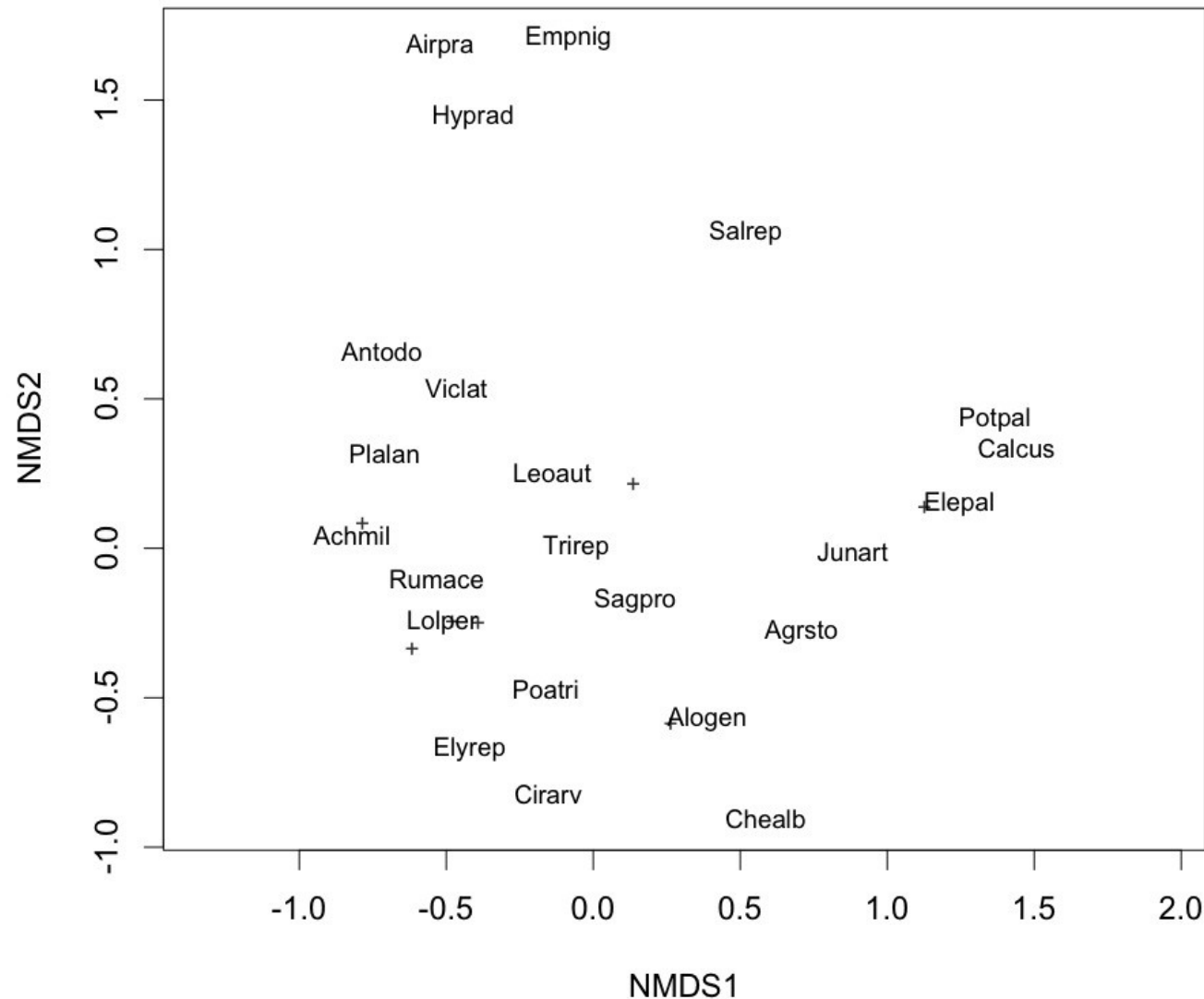
# Unconstrained ordination with NMDS





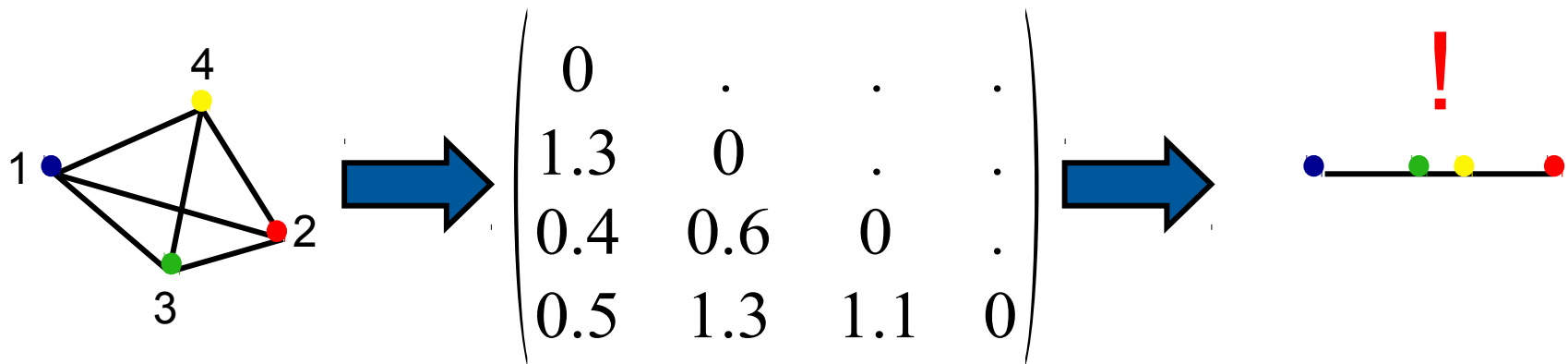
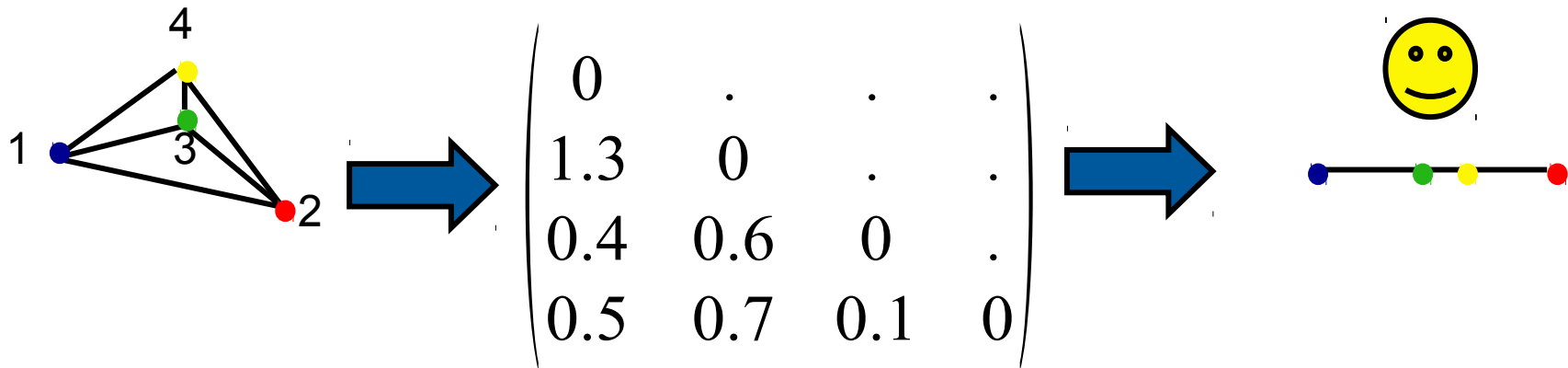
# Non-metric multidimensional scaling

- Unconstrained ordination for different distance metrics, based on ordered distances
- Suitable for ecological data
- Not based on eigenvalues, no partitioning of variance
- Very robust and flexible



# Understanding NMDS

The challenge of visualising distances in a lower dimension:



NMDS does not preserve absolute distances between objects, only ordered/ranked distances  
→ Easier to reduce dimensionality

# Steps of NMDS algorithm

1. Determine distance matrix from raw data
2. Choose initial configuration (often based on MDS/PCoA) in lower dimensional space
3. Determine distance matrix for this configuration
4. Determine disparities using monotone regression and pool adjacent violators (PAV) algorithm
5. Find a new configuration with higher similarity to the initial distance matrix
6. Go to 3. (if fit does not improve on many iterations → 7.)
7. Evaluate goodness of fit of final configuration

# From distance to disparity matrix

Distance matrix for data

$$\Delta = \begin{pmatrix} 0 & 9 & 4 & 10 & 7 \\ 9 & 0 & 3 & 1 & 2 \\ 4 & 3 & 0 & 8 & 6 \\ 10 & 1 & 8 & 0 & 5 \\ 7 & 2 & 6 & 5 & 0 \end{pmatrix}$$

Ordered distances of distance matrix

$$\delta_{24} < \delta_{25} < \delta_{23} < \delta_{13} < \delta_{45} < \delta_{35} < \delta_{15} < \delta_{34} < \delta_{12} < \delta_{14}$$

Distance matrix of the initial configuration

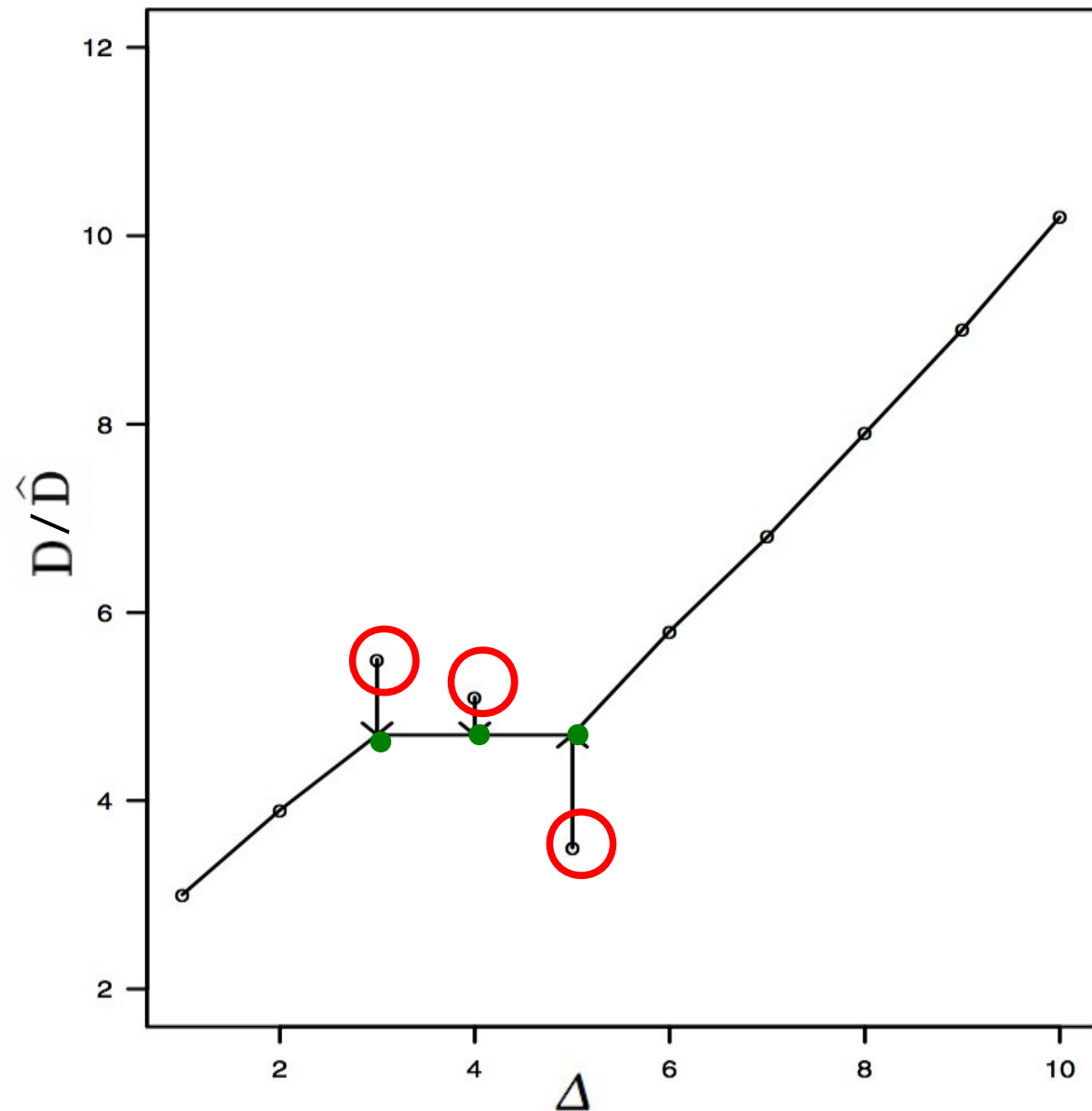
$$\mathbf{D} = \begin{pmatrix} 0 & 9.0 & 5.1 & 10.2 & 6.8 \\ 9.0 & 0 & 5.5 & 3.0 & 3.9 \\ 5.1 & 5.5 & 0 & 7.9 & 5.8 \\ 10.2 & 3.0 & 7.9 & 0 & 3.5 \\ 6.8 & 3.9 & 5.8 & 3.5 & 0 \end{pmatrix}$$

# Monotone regression

$$\Delta = \begin{pmatrix} 0 & 9 & 4 & 10 & 7 \\ 9 & 0 & 3 & 1 & 2 \\ 4 & 3 & 0 & 8 & 6 \\ 10 & 1 & 8 & 0 & 5 \\ 7 & 2 & 6 & 5 & 0 \end{pmatrix}$$

$$\mathbf{D} = \begin{pmatrix} 0 & 9.0 & 5.1 & 10.2 & 6.8 \\ 9.0 & 0 & 5.5 & 3.0 & 3.9 \\ 5.1 & 5.5 & 0 & 7.9 & 5.8 \\ 10.2 & 3.0 & 7.9 & 0 & 3.5 \\ 6.8 & 3.9 & 5.8 & 3.5 & 0 \end{pmatrix}$$

$$\hat{\mathbf{D}} = \begin{pmatrix} 0 & 9.0 & 4.7 & 10.2 & 6.8 \\ 9.0 & 0 & 4.7 & 3.0 & 3.9 \\ 4.7 & 4.7 & 0 & 7.9 & 5.8 \\ 10.2 & 3.0 & 7.9 & 0 & 4.7 \\ 6.8 & 3.9 & 5.8 & 4.7 & 0 \end{pmatrix}$$



# From distance to disparity matrix

Distance matrix for data

$$\Delta = \begin{pmatrix} 0 & 9 & 4 & 10 & 7 \\ 9 & 0 & 3 & 1 & 2 \\ 4 & 3 & 0 & 8 & 6 \\ 10 & 1 & 8 & 0 & 5 \\ 7 & 2 & 6 & 5 & 0 \end{pmatrix}$$

Ordered distances of distance matrix

$$\delta_{24} < \delta_{25} < \delta_{23} < \delta_{13} < \delta_{45} < \delta_{35} < \delta_{15} < \delta_{34} < \delta_{12} < \delta_{14}$$

Ordered distances of disparity matrix

$$\hat{d}_{24} \leq \hat{d}_{25} \leq \hat{d}_{23} \leq \hat{d}_{13} \leq \hat{d}_{45} \leq \hat{d}_{35} \leq \hat{d}_{15} \leq \hat{d}_{34} \leq \hat{d}_{12} \leq \hat{d}_{14}$$

Disparity matrix

$$\hat{\mathbf{D}} = \begin{pmatrix} 0 & 9.0 & 4.7 & 10.2 & 6.8 \\ 9.0 & 0 & 4.7 & 3.0 & 3.9 \\ 4.7 & 4.7 & 0 & 7.9 & 5.8 \\ 10.2 & 3.0 & 7.9 & 0 & 4.7 \\ 6.8 & 3.9 & 5.8 & 4.7 & 0 \end{pmatrix}$$

# Goodness of fit for NMDS

$$\text{STRESS 1} = \sqrt{\frac{\sum_{i < j} (d_{i,j} - \widehat{d}_{i,j})^2}{\sum_{i < j} d_{i,j}^2}}$$

Value of <b>STRESS1</b>	Goodness of configuration
$< 0.05$	excellent
$< 0.10$	good
$< 0.15$	medium
$> 0.15$	bad

## Implementation of NMDS in R

monoMDS() {vegan}	Basic function for NMDS
metaMDS() {vegan}	„Shotgun“ method
cmdscale() {stats}	(Metric) multidimensional scaling
cmds() {mclust}	

# What does the “shotgun” method do?

`metaMDS()` {vegan}

1. Data transformation (Square-root and Wisconsin double transformation)
2. Calculation of distance matrix based on the selected similarity coefficient (defaults to Bray-Curtis)
3. Adjustment if no shared occurrences of species
4. Several random starts for initial configuration
5. Centring and rotation of ordination (highest dispersion on 1<sup>st</sup> axis)
6. Scaling (1 unit means halving of community similarity)
7. Calculation of species scores as weighted averages of sites



# Limitations of NMDS

- Results dependent on initial configuration
- Loss of information due to ordered rank ordination
  - Information on absolute distances lost
  - No partitioning of variance
- Interpretation difficult if more than 2 or 3 dimensions required (i.e. to yield low STRESS1 value)
- Significant fit of environmental variables to ordered distances more difficult to interpret than for unconstrained variance-based methods

# Tools for complex data analysis

University of Koblenz-Landau 2018/19



Ralf B. Schäfer

These slides and notes complement the lecture with exercises “Tools for complex data analysis” for ecotoxicologists and environmental scientists. Do not hesitate to contact me if you have any comments or you find any errors (slides, slide notes, or code): [schaefer-ralf@uni-landau.de](mailto:schaefer-ralf@uni-landau.de)

While I made notes below the slides, some aspects are only mentioned in the R demonstration associated with the lecture.

# **RDA, similarity measures and NMDS**

## **Contents**

- 1. Learning targets, constrained ordination  
and RDA**
2. Diagnosis and assumptions of RDA,  
extensions and guidance for method  
selection
3. Similarity and distance measures
4. Non-metric multidimensional scaling (NMDS)

# Learning targets

- Understanding the basics of RDA.
- Knowledge on the calculation of commonly used association measures.
- Understanding their suitability for ecological data.
- Understanding the mathematical background and how to conduct a NMDS.

# Learning targets and study questions

- Understanding the basics of RDA.
  - How many constrained axes has an RDA and how are they related to the descriptors?
  - How does scaling influence the interpretation of a triplot?
- Knowledge on the calculation of commonly used association measures.
  - Which association is measured with similarity measures?
  - Outline the calculation of the Bray-Curtis and the Jaccard coefficient.
- Understanding their suitability for ecological data.
  - Explain the double-zero problem.
  - What is the species abundance paradox?

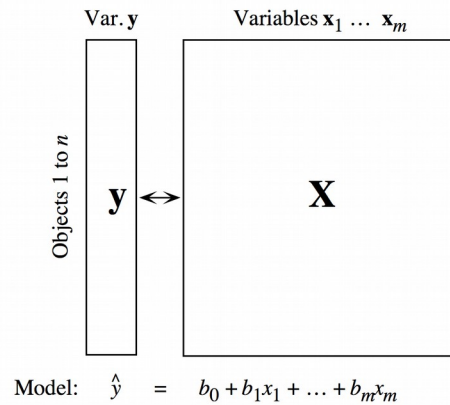
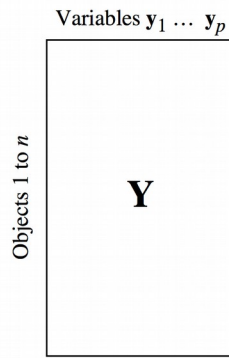
# Learning targets and study questions

- Understanding the mathematical background and how to conduct a NMDS.
  - What are the main differences between NMDS and PCA?
  - Which three matrices are computed during NMDS?
  - Outline the major elements of the algorithm used to compute the NMDS.
  - Discuss limitations of NMDS.

# Constrained ordination methods

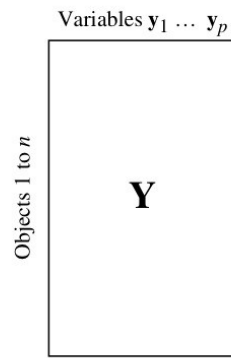
(a) Simple ordination of matrix **Y**:  
principal comp. analysis (PCA)  
correspondence analysis (CA)

(b) Ordination of **y** (single axis) under  
constraint of **X**: multiple regression

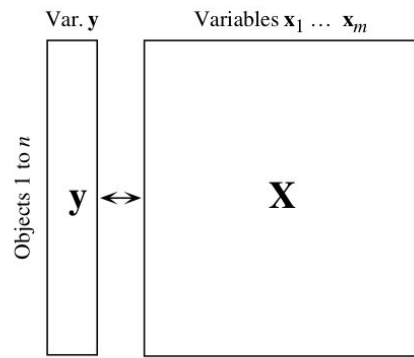


# Constrained ordination methods

(a) Simple ordination of matrix **Y**:  
principal comp. analysis (PCA)  
correspondence analysis (CA)

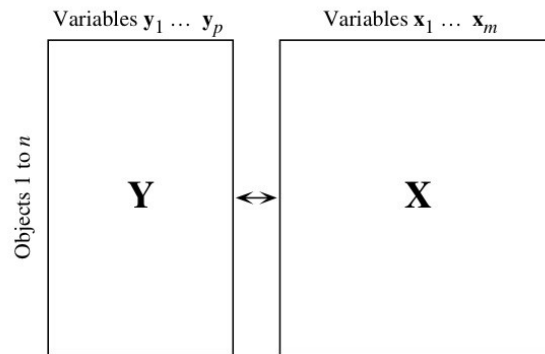


(b) Ordination of **y** (single axis) under  
constraint of **X**: multiple regression



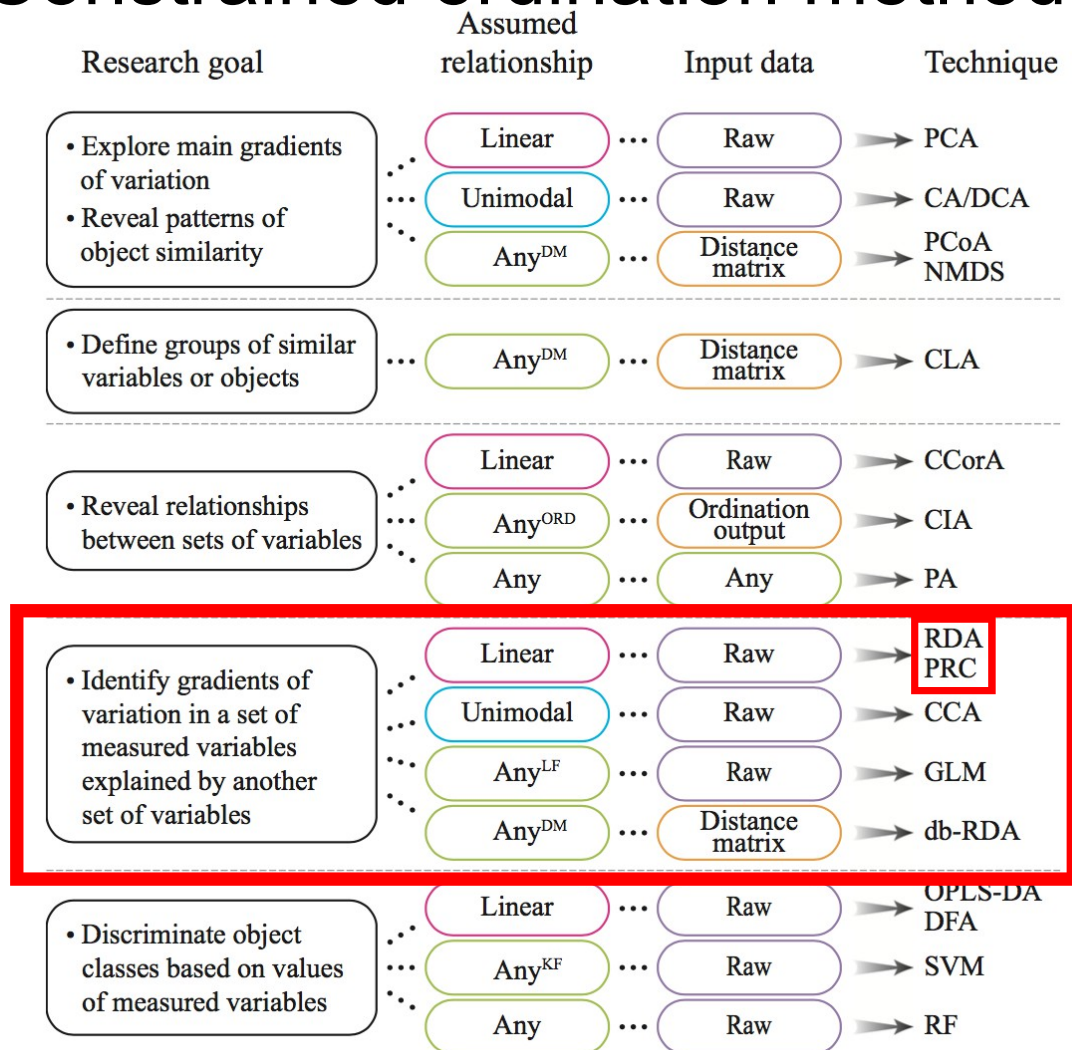
Model:  $\hat{y} = b_0 + b_1x_1 + \dots + b_mx_m$

(c) Ordination of **Y** under constraint of **X**:  
redundancy analysis (RDA)  
canonical correspondence analysis (CCA)





# Constrained ordination methods



8

Paliy & Shankar 2016 *Mol Ecol*:1032

We will discuss RDA in detail, PRC is explained in extra course materials. CCA and db-RDA will also be briefly discussed. An alternative approach represent multivariate GLMs that extend the framework for one response to multiple response variables. Details on this method and its advantages compared to ordination methods are given in Wang et al. 2012 and Warton et al. 2011, a tutorial for the application and a comparison to PRCs is given in Szöcs et al. 2015.

Paliy O. & Shankar V. (2016) Application of multivariate statistical techniques in microbial ecology. *Molecular Ecology* 25, 1032–1057.

Szöcs E. et al. (2015) Analysing chemical-induced changes in macroinvertebrate communities in aquatic mesocosm experiments: a comparison of methods. *Ecotoxicology* 24, 760–769.

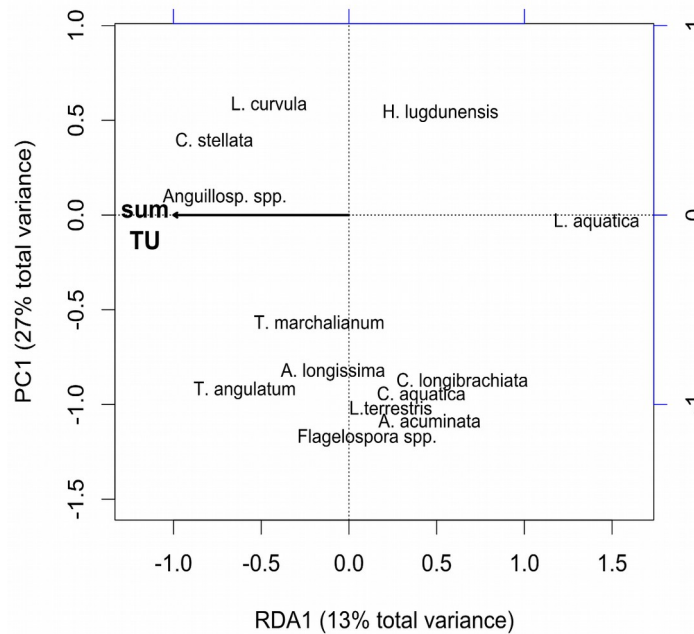
Wang Y., Naumann U., Wright S.T. & Warton D.I. (2012) mvabund – an R package for model-based analysis of multivariate abundance data. *Methods in Ecology and Evolution* 3, 471–474.

Warton D.I., Wright S.T. & Wang Y. (2011) Distance-based multivariate analyses confound location and dispersion effects. *Methods in Ecology and Evolution* 3, 89–101.

# Redundancy Analysis (RDA)

**Aim:** Display and explain variation in set of response variables constrained by second set of predictor variables  
→ Links multivariate multiple regression and PCA

**Example:** Which variable(s) do best explain the variation in fungal communities sampled along a gradient of fungicide toxicity?




Redundancy = explained variance

# Mathematical background of RDA

**Aim:** Display and explain variation in set of response variables constrained by second set of predictor variables  
→ Links multivariate multiple regression and PCA

Remember: Multiple linear regression in matrix form

$$\begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{pmatrix} \quad \Rightarrow \quad \hat{y} = \mathbf{X}b$$



$$b = (\mathbf{X}^T \mathbf{X})^{-1}(\mathbf{X}^T y)$$

Substitution yields:  $\hat{y} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1}(\mathbf{X}^T y)$

Reformulation for multivariate multiple regression with several  $y$ :

$$\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1}(\mathbf{X}^T \mathbf{Y})$$

10

Note that in previous lectures we used the matrix notation for  $y$  in the equations, i.e.  $\mathbf{Y}$  where  $\mathbf{Y}$  is a  $n \times 1$  matrix. Indeed, we can express the response both as vector or as  $n \times 1$  matrix. We used the latter in the context of matrix algebra.

Here, we want to emphasise the transition from a response vector  $y$  to a response matrix  $\mathbf{Y}$ .

# Mathematical background of RDA

$$\hat{Y} = X(X^T X)^{-1}(X^T Y)$$

RDA uses variance-covariance matrix of  $\hat{Y} \Rightarrow \Sigma_{Y^T Y}$

Usually, this is not known and the sample variance-covariance matrix (also called Dispersion matrix) is estimated from the observations:

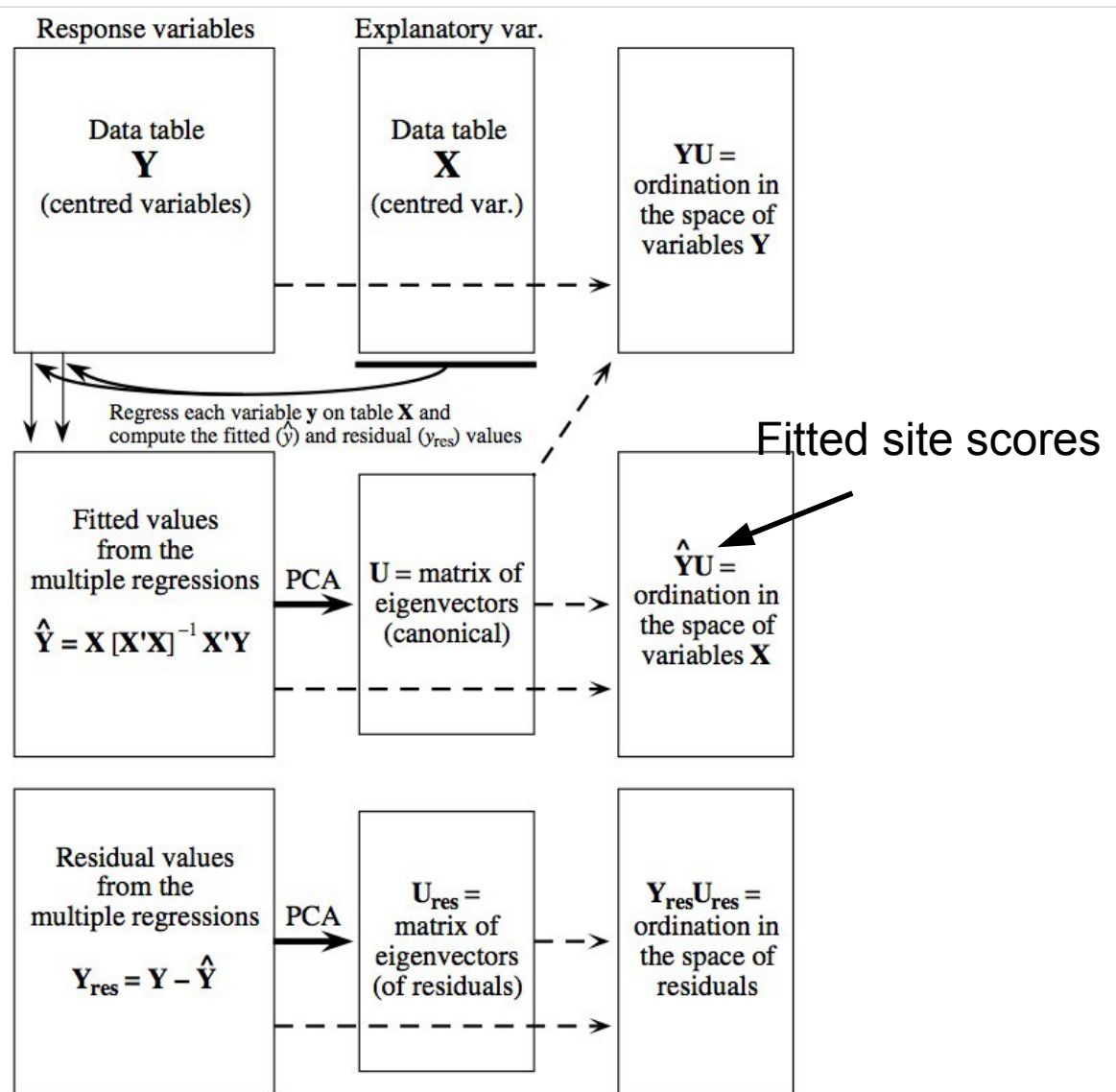
$$S_{\hat{Y}^T \hat{Y}} = \frac{1}{n-1} \hat{Y}^T \hat{Y}$$

and used in a PCA:  $S_{\hat{Y}^T \hat{Y}} a = \lambda a$   Eigenvector

Eigenvalue problem



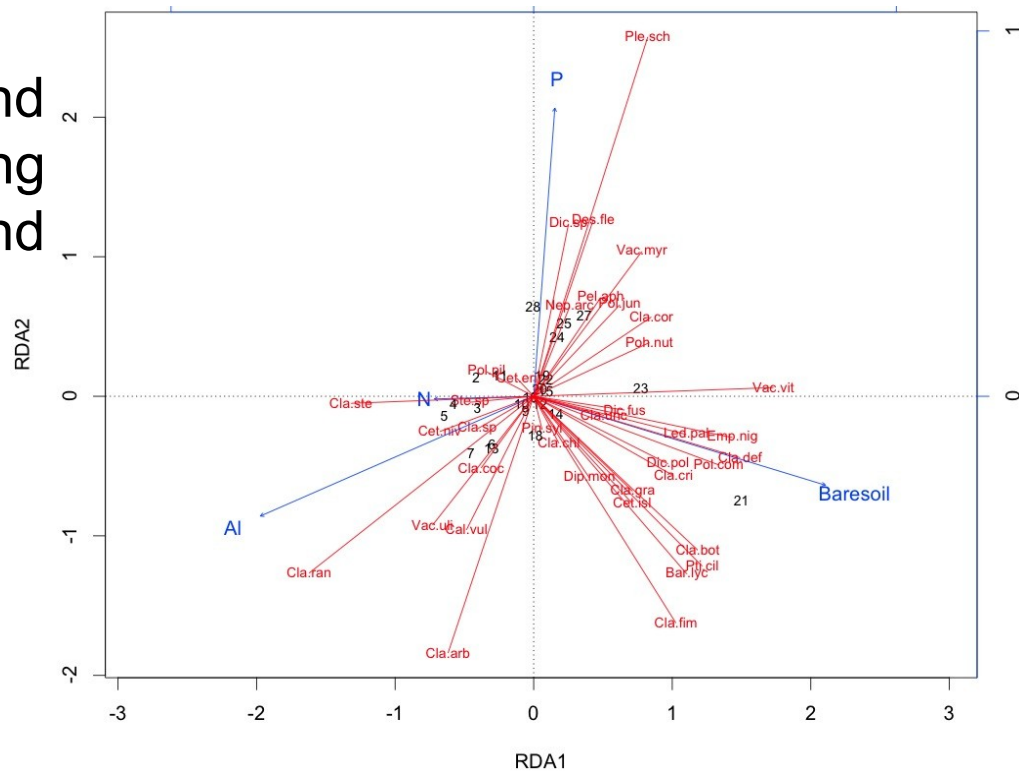
Eigenvectors linear combinations of predictors



For details on the algebra behind RDA refer to Legendre & Legendre 2012: 637ff

# RDA results

- Triplot with relationship between species, sites and env. variables
- Eigenvalues and variance partitioning (constrained and unconstrained)
- Site scores
- Species scores
- Biplot scores for variables



13

We will discuss the RDA results in detail during the demonstration in R.

# RDA, similarity measures and NMDS

## Contents

1. Learning targets, constrained ordination and RDA
- 2. Diagnosis and assumptions of RDA, extensions and guidance for method selection**
3. Similarity and distance measures
4. Non-metric multidimensional scaling (NMDS)

# RDA axes and variable importance

How many RDA axes are required?

- Hypothesis test (permutation-based) recommended (Legendre et al. *Methods Ecol. Evol.* 2011)

How many environmental variables are needed and how important are they?

- Manual and automatic model-building with *adj. R<sup>2</sup>* as goodness of fit criteria (as for multiple linear regression)
- Variance partitioning between different models to determine explained variance of individual variables

Legendre P., Oksanen J. & ter Braak C.J.F. (2011) Testing the significance of canonical axes in redundancy analysis. *Methods in Ecology and Evolution* 2, 269–277.



# Assumptions and extensions of RDA

- Independence of observations (sites)
- Linear relationship between explanatory and response variables → see next slide
- No multicollinearity between explanatory variables
- $n$  (sites)  $\gg p$  (predictors) to reliably infer  $p$  importance
- RDA can be employed for multivariate ANOVA (see Borcard et al. 2011: 185 ff)
- RDA over time important for ecotoxicological experiments:  
→ Principal Response Curves (PRC) that deliver time-dependent treatment effects relative to control (van den Brink & ter Braak 1999 *ET&C* 18 (2): 138-148)

16

The data preparation steps of RDA include similar steps as for multiple linear regression analysis (e.g. checking for multicollinearity, distribution of variables). To improve interpretation and to increase the strength of the relationship between predictors and organisms (i.e. the explained variance on first few RDA axes), the rarest species are sometimes removed. Legendre & Birks (2012) mention the suggestion of Daniel Borcard to remove rare species until this shows no effect on the first few RDA axes.

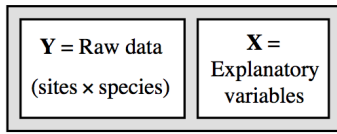
PRC is described in more detail in extra materials for this lecture that can be accessed here: [https://github.com/EDiLD/permanova\\_lecture/tree/master/prc](https://github.com/EDiLD/permanova_lecture/tree/master/prc)  
Note that PRC is highly relevant for students of Ecotoxicology

Legendre P. & Birks H.J.B. (2012) From Classical to Canonical Ordination. In: Tracking Environmental Change Using Lake Sediments: Data Handling and Numerical Techniques. (Eds H.J.B. Birks, A.F. Lotter, S. Juggins & J.P. Smol), pp. 201–248. Springer Netherlands, Dordrecht.

# RDA approaches

## How to assess gradient length?

(a) Classical approach: RDA preserves the Euclidean distance, CCA preserves the chi-square distance

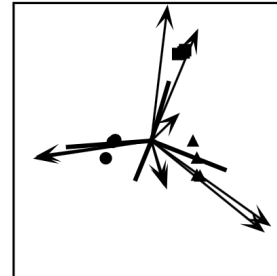


Short gradients: CCA or RDA

Long gradients: CCA

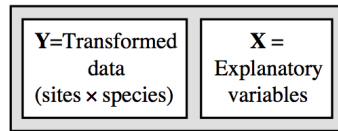
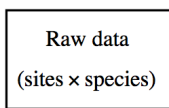
- test for higher order terms (Borcard et al. 2011: 190ff)
- Axis length in DCA

Canonical ordination triplot



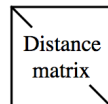
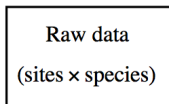
Representation of elements:  
Species = arrows  
Sites = symbols  
Explanatory variables = lines

(b) Transformation-based RDA (tb-RDA) approach: preserves a distance obtained by data transformation

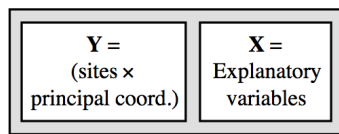


RDA

(c) Distance-based RDA (db-RDA) approach: preserves a pre-computed distance



PCoA



RDA

## DCA: Detrended Correspondence Analysis

Transformation-based RDA is discussed in:

Legendre P. & Gallagher E.D. (2001) Ecologically meaningful transformations for ordination of species data. *Oecologia* 129, 271–280.

Information on db-RDA can be found in the following papers:

Legendre, P. & Anderson, M. J. (1999) Distance-based redundancy analysis: Testing multispecies responses in multifactorial ecological experiments. *Ecol. Monogr.* 69 (1), 1-24

McArdle, B. H. & Anderson, M. J. (2002) Fitting multivariate models to community data: A comment on distance-based redundancy analysis. *Ecology* 82 (1), 290-297.

An example for a study where db-RDA is used:

Szöcs, E., Kefford, B. J. & Schäfer, R. B. (2012) Is there an interaction of the effects of salinity and pesticides on the community structure of 17 macroinvertebrates? *Sci. Total Environ.* 437 (1), 121-126.

# Further constrained ordination methods

## Canonical Correspondence Analysis (CCA)

- Widely used
- Extension of (unconstrained) correspondence analysis
- Similar to RDA, but assumes unimodal distribution ( $\chi^2$ -distance) of species along environmental gradient
- In R: model building as for RDA `cca()` {vegan}

## Constrained additive Ordination (CAO)

- Comparatively new
- derives response of each species to main environmental gradient from data → no linear or unimodal model assumed
- mixture of Generalized Additive Models (GAMs) and Canonical Gaussian Ordination
- computationally demanding `cao()` {VGAM}
- In R: implemented in extra package

18

We touch only briefly on CCA and CAO. A short description with mathematical background of CCA is given in Legendre & Legendre (2012) and Zuur et al. (2007). A very readable introduction to CCA can be found in Leps & Smilauer (2003), or on an advanced level in ter Braak & Verdonschot (1995). CCA is widely used in ecology because often an unimodal distribution is assumed or known. CCA is implemented in the R package *vegan*, which provides an introduction (section Constrained ordination): <https://cran.r-project.org/web/packages/vegan/vignettes/intro-vegan.pdf> (see also: Borcard et al (2018)).

CAO represents a novel and flexible ordination approach (Yee 2006, Yee 2015) that can be used for any species response to environmental gradients. The technique is implemented in R (package *VGAM*) and allows visualisation of individual species responses' to environmental gradients, as well as to conduct an ordination and extract main gradients. The method is computationally demanding (calculations can take up to several hours, depending on the model specification and size of the data set).

Several of the references can be found in the literature list.

Leps J. & Smilauer P. (2003) *Multivariate Analysis of Ecological Data using CANOCO*. University Press, Cambridge.

Ter Braak, C.J.F. & Verdonschot, P.F.M. (1995) Canonical correspondence analysis and related multivariate methods in aquatic ecology. *Aquatic Sciences*, 57, 255-289.

Yee, T. W. (2006). Constrained additive ordination. *Ecology* 87, 203-213.

Yee, T.W. (2015) Constrained additive ordination *in*: *Vector Generalized Linear and Additive Models*, Springer, DOI: 10.1007/978-1-4939-2818-7 7

# When to use what?

Numerical methods to *forecast* one or several descriptors (response or dependent variables) using other descriptors (explanatory or independent variables). In parentheses, identification of the section where a method is discussed.

- 
- 1) Forecasting the structure of a *single* descriptor, or *indirect comparison* ..... see 2
  - 2) The response variable is quantitative ..... see 3
  - 3) The explanatory variables are quantitative ..... see 4
    - 4) Null or low correlations among explanatory variables: *multiple linear regression* (10.3); *nonlinear regression* (10.3)
    - 4) High correlations among explanatory variables (collinearity): *ridge regression* (10.3); *regression on principal components* (10.3)
  - 3) The explanatory variables are qualitative: *dummy variable regression* (10.3)
  - 2) The response variable is qualitative (*or* a classification) ..... see 5
    - 5) Response: two or more groups; explanatory variables are quantitative (but qualitative variables may be recoded into dummy variables): *identification functions in discriminant analysis* (11.3)
    - 5) Response: binary (presence-absence); explanatory variables are quantitative (but qualitative variables may be recoded into dummy var.): *logistic regression* (10.3)
  - 2) The response and explanatory variables are quantitative, but they display a nonlinear relationship: *nonlinear regression* (10.3)
  - 1) Forecasting the structure of a *multivariate* data matrix ..... see 6
    - 6) *Direct comparison* ..... see 7
      - 7) Linear modelling: *redundancy analysis* (RDA, 11.1); *canonical correspondence analysis* (CCA, 11.2)
      - 7) Find a tree-like decision model: *multivariate regression tree analysis* (MRT, 8.11)
    - 6) *Indirect comparison* ..... see 8
      - 8) Ordination in reduced space: each axis is treated in the same way as a single quantitative descriptor ..... see 2
      - 8) Clustering: each partition is treated as a qualitative descriptor ..... see 2
- 

Some of the methods have already been discussed, others will be in the remainder of the course.





# RDA, similarity measures and NMDS

## Contents

1. Learning targets, constrained ordination and RDA
2. Diagnosis and assumptions of RDA, extensions and guidance for method selection
- 3. Similarity and distance measures**
4. Non-metric multidimensional scaling (NMDS)

# Measuring association

## Example: Species observations in 4 streams

Site				
1	0	400	0	0
2	0	0	10	0
3	2	280	3	3
4	12	60	80	50

## What is the relationship between 1) objects 2) descriptors?

- Relationship between objects (sites): distance or similarity measures
- Relationship between descriptors (species): Dependence measures (e.g. covariance or correlation between environmental variables)

21

Analyses of the association between objects are defined as *Q Mode*, analyses of the association between descriptors are defined as *R Mode* (see Legendre & Legendre 2012: 265-268 for details).





# Similarity measures: Presence-absence

## Simple matching coefficient





		Site 1		
		present	absent	
Site 2	present	a	b	a + b
	absent	c	d	c + d
Sum		a + c	b + d	

$$S_m = \frac{a + d}{a + b + c + d}$$

Exercise: Calculate  $S_m$  for the data below with and without the 1. and 4. species. How do these species influence  $S_m$ ?

Site				
1	0	400	0	0
2	0	0	10	0

# Similarity measures: Presence-absence

Site				
1	0	400	0	0
2	0	0	10	0

$$S_m = \frac{a+d}{a+b+c+d}$$

Calculation with all species:

$$a = 0, b = 1, c = 1, d = 2 \rightarrow S_m = 2/4 = 0.5$$

Calculation without species 1 and 4:

$$a = 0, b = 1, c = 1, d = 0 \rightarrow S_m = 0/2 = 0$$

Species absence influences similarity between sites.

Not desirable: joint absence of species does not indicate ecological similarity and number of joint absences is arbitrary

→ **Double-Zero problem**



# Widely used similarity measures

## Jaccard coefficient (=Jaccard similarity index)

		Site 1		
		present	absent	
Site 2	present	a	b	a + b
	absent	c	d	c + d
Sum		a + c	b + d	

$$S_j = \frac{a}{a+b+c}$$

- used for binary data
- ignores joint absences (d)

## Bray-Curtis coefficient

- used for abundance data
- range: 0 - 1 (if all  $x_k \geq 0$ )
- data transformation often required to reduce weight of dominant taxa

$$S_{BC}(i, j) = \frac{2 \sum_{k=1}^n \min(x_{i,k}, x_{j,k})}{\sum_{k=1}^n |x_{i,k} + x_{j,k}|}$$

24

$X_{i,k}$  and  $x_{j,k}$  is the abundance of taxon  $k$  for site  $i$  and  $j$





Similarity indices that ignore shared absences to cope with the double-zero problem are asymmetrical, whereas similarity indices that include  $d$  are symmetrical (since absence and presence are treated in the same manner).

The simple matching coefficient and the Jaccard coefficient both range from 0 to 1. The Jaccard dissimilarity  $D_j$  is given as:

$D_j = b/(a+b+c)$  or  $1-S_j$ .  $S_j$  is often used for presence-absence data and gives equal weight to all species (except for double-absences).

The Bray Curtis coefficient is more appropriately called Steinhaus coefficient (see Legendre & Legendre 2012: 311).

# Example: Bray-Curtis coefficient

Site				
1	0	400	5	0
2	0	0	10	0
Min	0	0	5	0
Sum	0	400	15	0

$$S_{BC}(i, j) = \frac{2 \sum_{k=1}^n \min(x_{i,k}, x_{j,k})}{\sum_{i=1}^n |x_{i,k} + x_{j,k}|}$$

## Calculation:

$$2*(0+0+5+0)/415 \rightarrow S_{BC} = 10/415 = 0.025$$

## Calculation after square-root transformation:

$$2*(0+0+5^{0.5}+0)/(400^{0.5}+5^{0.5}+10^{0.5}) \rightarrow S_{BC} = 0.18$$

## Calculation after double square-root transformation:

$$2*(0+0+5^{0.25}+0)/(400^{0.25}+5^{0.25}+10^{0.25}) \rightarrow S_{BC} = 0.39$$

25

The dissimilarity  $D_{BC}$  for Bray-Curtis is  $1-S_{BC}$  and can be calculated according to:

$$D_{BC}(i, j) = \frac{\sum_{k=1}^n |x_{i,k} - x_{j,k}|}{\sum_{k=1}^n |x_{i,k} + x_{j,k}|}$$

In the extreme case of samples without any species – perhaps due to a chemical spill – a dummy species can be added to the data to enable calculation of the Bray-Curtis coefficient (Clarke et al 2006).

Note that the  $\log(x+1)$  transformation increases the weight of rare taxa, whereas square-root or double-square root transformation does not. For a discussion on transformations for ecological data see Legendre & Gallagher (2001).

Clarke, K.R; Somerfield, P.J; Chapman, M.G (2006): On resemblance measures for ecological studies, including taxonomic dissimilarities and a zero-adjusted Bray-Curtis coefficient for denuded assemblages. *J. Experim. Mar. Biol. Ecol.*, 330, 55–80.

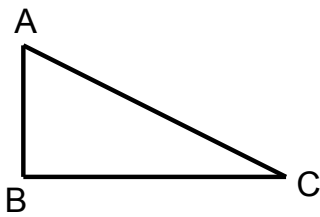
Legendre, P. & Gallagher, E.D. (2001) Ecologically meaningful transformations for ordination of species data. *Oecologia*, 129, 271-280.

25

# Distance measures

## Association measures meeting triangle inequality criterion

(following Everitt et al. 2011 *Cluster Analysis*. John Wiley & Sons: 49)



### Triangle inequality criterion

$d(A,B) + d(B,C) \geq d(A,C)$ , where  $d$  is distance function

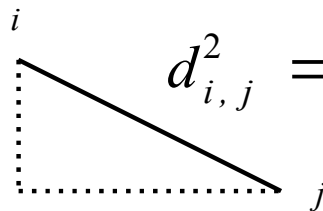
Sum of any two sides of triangle always  $\geq$  third side

Important for geometrical representation (e.g. ordination)

## Euclidean distance: Most frequently used distance measure

$$d_{i,j} = \sqrt{\sum_{k=1}^n (x_{i,k} - x_{j,k})^2}$$

### Two dimensional case:



$$d_{i,j}^2 = (x_{i,1} - x_{j,1})^2 + (x_{i,2} - x_{j,2})^2$$

26

Note that Legendre & Legendre (2012) do not reserve the term distance measure for dissimilarity measures that meet the triangle inequality criterion.

Several measures can be square-root transformed to meet the triangle inequality criterion, see Legendre & Legendre 2012: 295-297 for details.

We have already learnt that the Mahalanobis distance can be used to evaluate the distance between two multivariate vectors (objects when following the terminology in the context of association measures) or between individual multivariate vectors (objects) and the overall mean vector, taking the covariance/correlation among the descriptors (here: species) into account, which renders the Mahalanobis distance independent from different scales in the data set. However, if the covariance matrix for two multivariate vectors is the identity matrix, then the Mahalanobis and Euclidean distance are equivalent. The Euclidean distance is sensitive to outliers and depends on the scales of the descriptors (species). Hence, for species data, the measure is typically dominated by the species with the largest absolute difference in abundance, unless the magnitude of absolute differences are similar between all species.

# Species abundance paradox

**Species x Site matrix**

Sites	Species		
	$y_1$	$y_2$	$y_3$
$x_1$	0	1	1
$x_2$	1	0	0
$x_3$	0	4	4

Euclidean  
distance



**Distance matrix**

Sites	Sites		
	$x_1$	$x_2$	$x_3$
$x_1$	0	1.732	4.243
$x_2$	1.732	0	5.745
$x_3$	4.243	5.745	0

Sites  $x_1$  and  $x_2$  share no species, but have a smaller distance than sites sharing species ( $x_1$  and  $x_3$ ).

→ Euclidean distance problematic for ecological data

# How to select a measure

- Many more association measures

(see Legendre & Legendre 2012: Chapter 7)

- Check literature of scientific field

- Refer to key in Legendre & Legendre 2012: 325-328

Choice of an association measure among objects (Q mode), to be used with chemical, geological physical, etc. descriptors (symmetrical coefficients, using double-zeros).

---

1) Association measured between individual objects	see 2
2) Descriptors: presence-absence or multistate (no partial similarities computed between states)	see 3
3) Metric coefficients: <i>simple matching</i> ( $S_1$ ) and derived coefficients ( $S_2, S_6$ )	
3) Semimetric coefficients: $S_3, S_5$	
3) Nonmetric coefficient: $S_4$	
2) Descriptors: multistate (states defined in such a way that partial similarities can be computed between them)	see 4
4) Descriptors: quantitative and dimensionally homogeneous	see 5
5) Differences enhanced by squaring: <i>Euclidean distance</i> ( $D_1$ ) and <i>average distance</i> ( $D_2$ )	

28

It should be obvious that no single measure fits all purposes, the selection should always be guided by the research question.

# Association measures in R

Function	Package	No. of measures	Weighing possible?
dist	stats	6	No
daisy	cluster	3	Yes
dsvdis	labdsv	7	Yes
vegdist	vegan	14 (easily expandable)	No
distance	ecodist	10 (easily expandable)	Yes
dist.*	ade4	~25 (in different functions)	Yes

# RDA, similarity measures and NMDS

## Contents

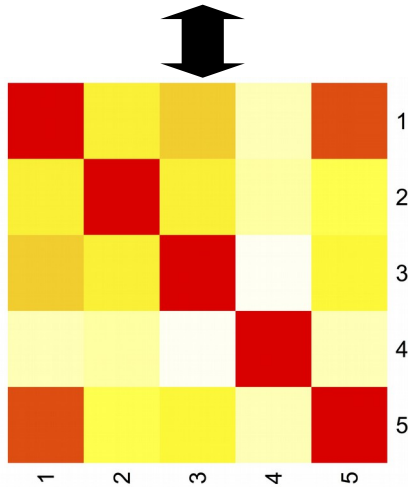
1. Learning targets, constrained ordination and RDA
2. Diagnosis and assumptions of RDA, extensions and guidance for method selection
3. Similarity and distance measures
- 4. Non-metric multidimensional scaling (NMDS)**

# Visualization of association measures

## Heatmap

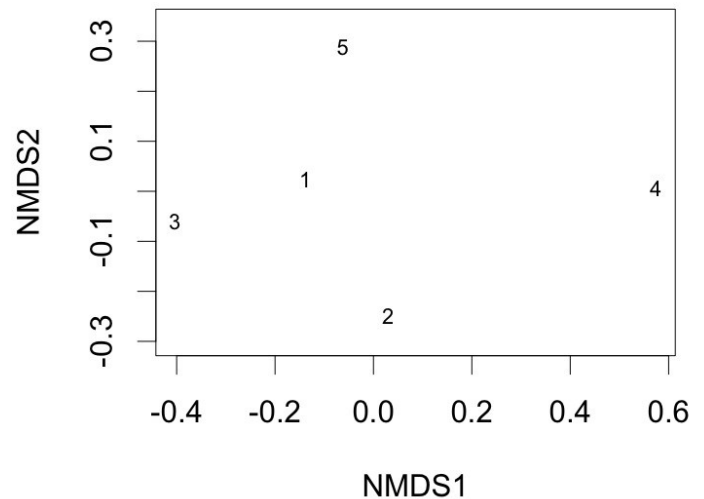
- Associations converted to colours
- Relationship easier to grasp

	1	2	3	4	5
1	0.00	0.69	0.60	0.92	0.22
2	0.69	0.00	0.70	0.89	0.80
3	0.60	0.70	0.00	0.98	0.72
4	0.92	0.89	0.98	0.00	0.92
5	0.22	0.80	0.72	0.92	0.00



## Ordination

- Works for measures that meet triangle inequality criterion (otherwise no clear geometrical interpretation possible)



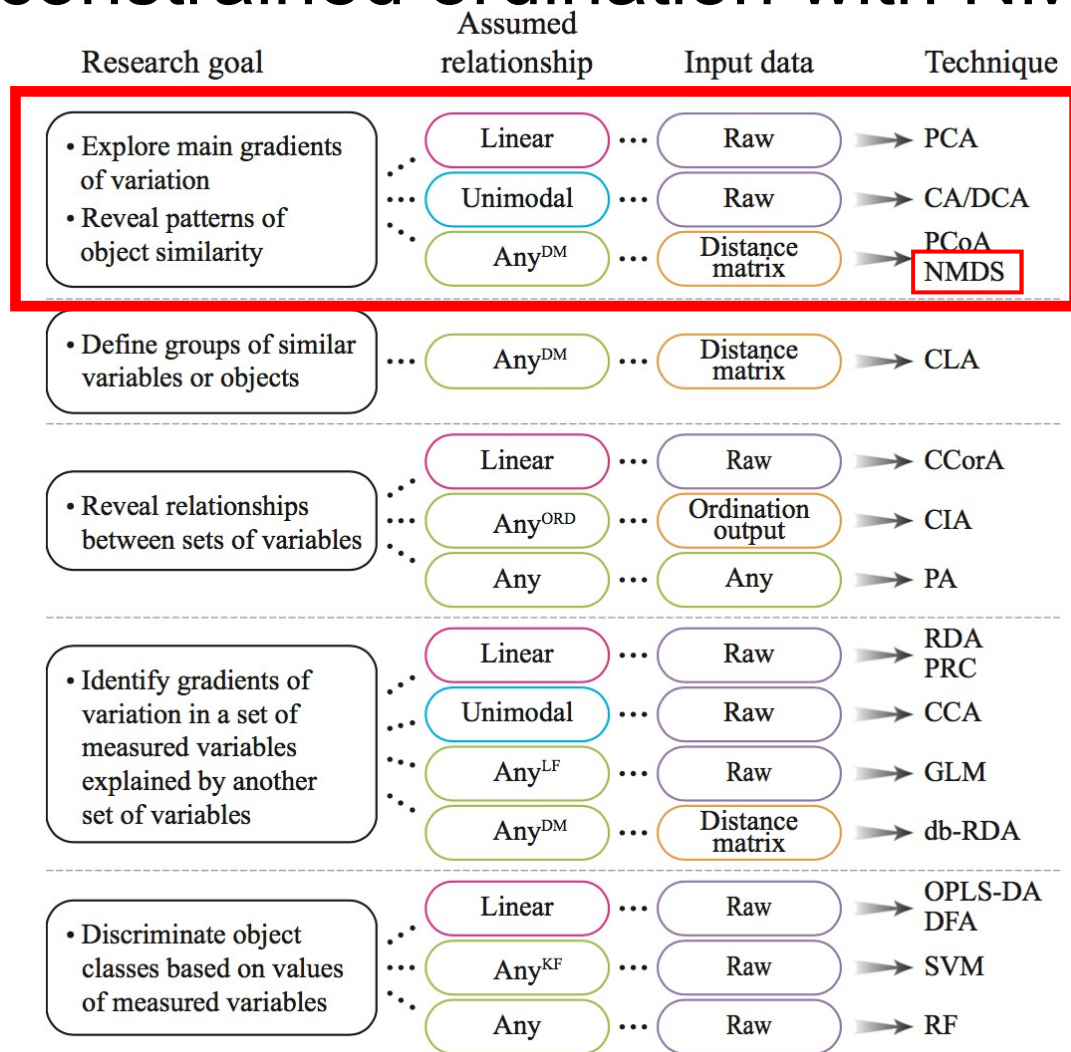
31

For an overview of techniques to visualize multivariate ecological data see:

Warton (2008). Raw data graphing: an informative but under-utilized tool for the analysis of multivariate abundances. *Austral. Ecol.* (33), 290–300.



# Unconstrained ordination with NMDS



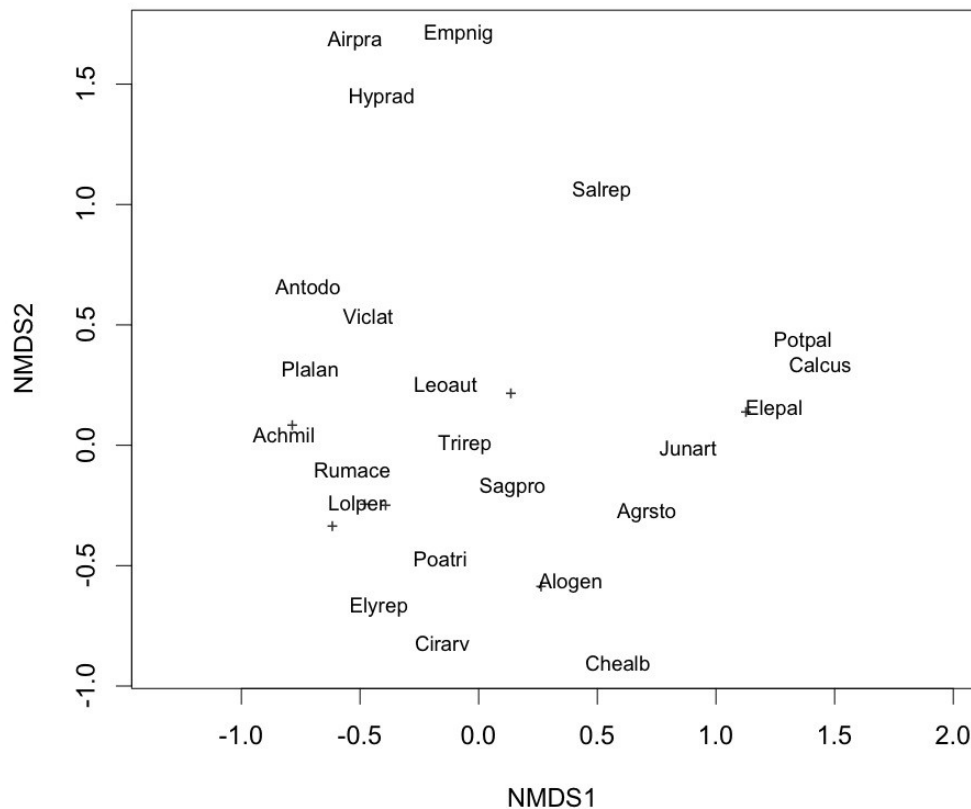
32

Paliy & Shankar 2016 *Mol Ecol*:1032

Paliy O. & Shankar V. (2016) Application of multivariate statistical techniques in microbial ecology. *Molecular Ecology* 25, 1032–1057.

# Non-metric multidimensional scaling

- Unconstrained ordination for different distance metrics, based on ordered distances
- Suitable for ecological data
- Not based on eigenvalues, no partitioning of variance
- Very robust and flexible



33

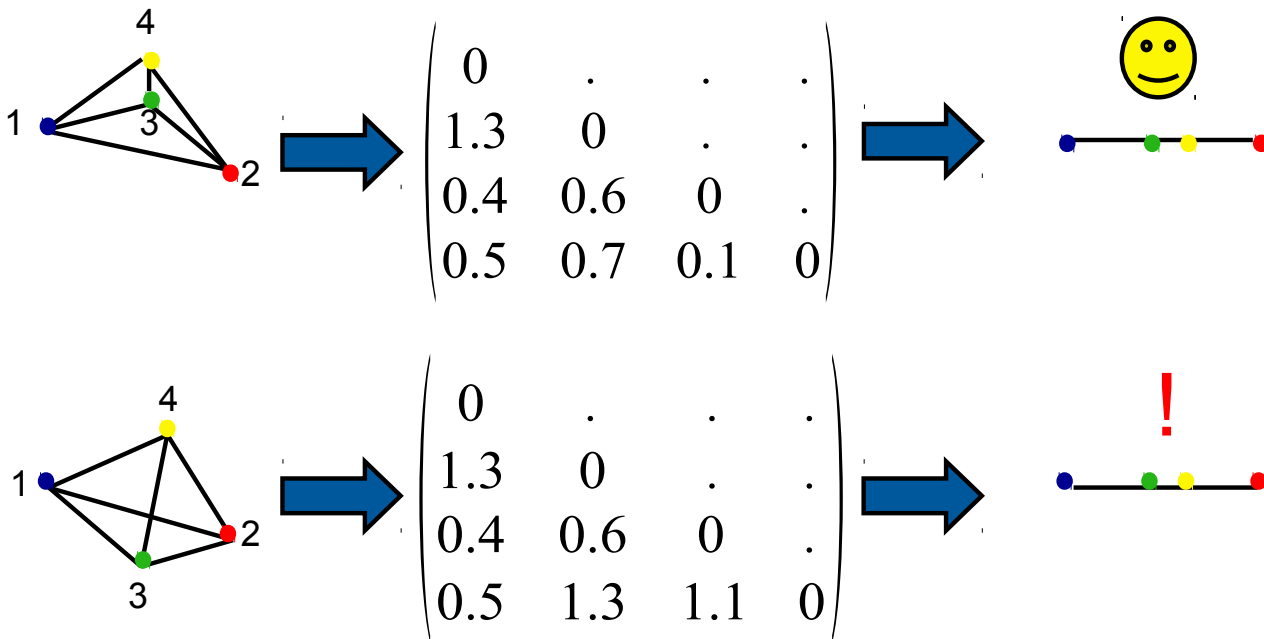
NMDS is the non-metric version of multidimensional scaling (MDS), which is also called Principal Coordinate Analysis (PCoA). PCoA preserves the original distances between objects, whereas NMDS intentionally relaxes this condition and only preserves the order of the distances. NMDS is therefore more robust to outliers and an appropriate two-dimensional representation of distances, though a distortion of original distances, can be easier obtained by NMDS than PCoA.

NMDS is based on distance matrices and circumvents the critical assumption of a linear gradient by only requiring monotonicity for the transformation of the distance matrix to a lower-dimensional space.

NMDS is regarded as one of the most robust unconstrained ordination methods (Minchin, P.R. 1987 An Evaluation of the Relative Robustness of Techniques for Ecological Ordination. *Vegetatio*, **69**, 89-107). It is very flexible as it can be used with a variety of distance measures.

# Understanding NMDS

The challenge of visualising distances in a lower dimension:



NMDS does not preserve absolute distances between objects, only ordered/ranked distances  
→ Easier to reduce dimensionality

# Steps of NMDS algorithm

1. Determine distance matrix from raw data
2. Choose initial configuration (often based on MDS/PCoA) in lower dimensional space
3. Determine distance matrix for this configuration
4. Determine disparities using monotone regression and pool adjacent violators (PAV) algorithm
5. Find a new configuration with higher similarity to the initial distance matrix
6. Go to 3. (if fit does not improve on many iterations → 7.)
7. Evaluate goodness of fit of final configuration

35

MDS: Multidimensional Scaling

PCoA: Principal Coordinate Analysis

MDS and PCoA are alternative terms for the same method.

The PAV algorithm searches for points that violate the monotonicity assumption and averages the distances of these points. See next slides and Härdle & Simar (2015).

Härdle W. & Simar L. (2015) Applied multivariate statistical analysis, Fourth Edition. Springer, Berlin Heidelberg New York Dordrecht London.

# From distance to disparity matrix

Distance matrix for data

$$\Delta = \begin{pmatrix} 0 & 9 & 4 & 10 & 7 \\ 9 & 0 & 3 & 1 & 2 \\ 4 & 3 & 0 & 8 & 6 \\ 10 & 1 & 8 & 0 & 5 \\ 7 & 2 & 6 & 5 & 0 \end{pmatrix}$$

Ordered distances of distance matrix

$$\delta_{24} < \delta_{25} < \delta_{23} < \delta_{13} < \delta_{45} < \delta_{35} < \delta_{15} < \delta_{34} < \delta_{12} < \delta_{14}$$

Distance matrix of the initial configuration

$$\mathbf{D} = \begin{pmatrix} 0 & 9.0 & 5.1 & 10.2 & 6.8 \\ 9.0 & 0 & 5.5 & 3.0 & 3.9 \\ 5.1 & 5.5 & 0 & 7.9 & 5.8 \\ 10.2 & 3.0 & 7.9 & 0 & 3.5 \\ 6.8 & 3.9 & 5.8 & 3.5 & 0 \end{pmatrix}$$

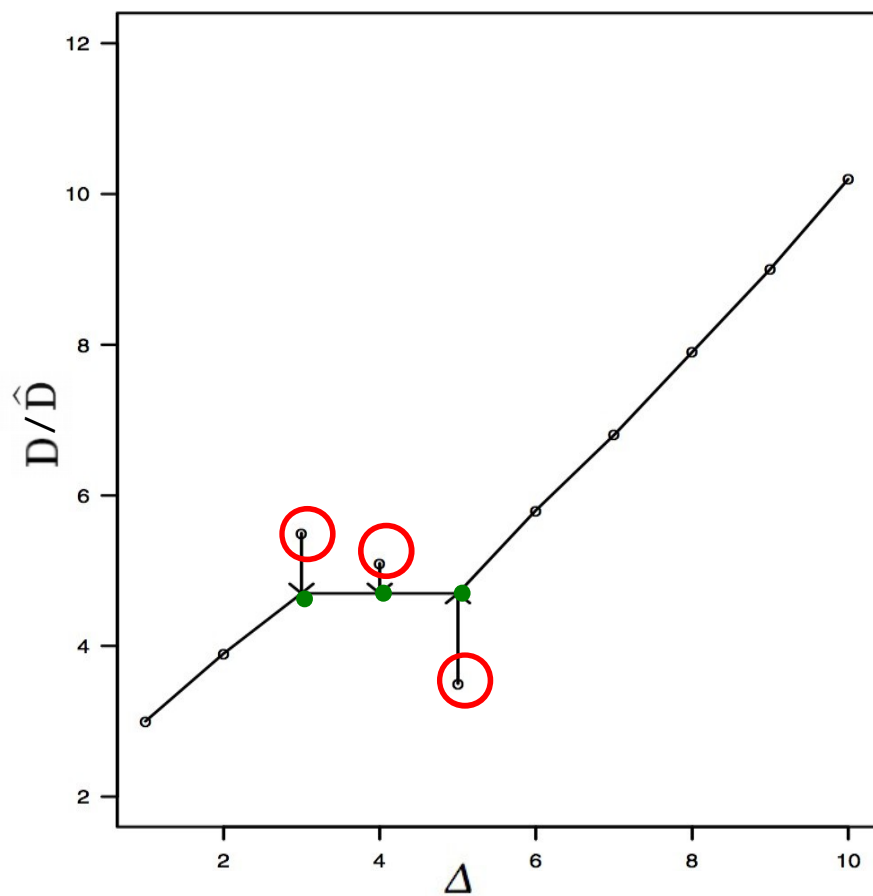
Example adopted from Handl (2010).

# Monotone regression

$$\Delta = \begin{pmatrix} 0 & 9 & 4 & 10 & 7 \\ 9 & 0 & 3 & 1 & 2 \\ 4 & 3 & 0 & 8 & 6 \\ 10 & 1 & 8 & 0 & 5 \\ 7 & 2 & 6 & 5 & 0 \end{pmatrix}$$

$$\mathbf{D} = \begin{pmatrix} 0 & 9.0 & 5.1 & 10.2 & 6.8 \\ 9.0 & 0 & 5.5 & 3.0 & 3.9 \\ 5.1 & 5.5 & 0 & 7.9 & 5.8 \\ 10.2 & 3.0 & 7.9 & 0 & 3.5 \\ 6.8 & 3.9 & 5.8 & 3.5 & 0 \end{pmatrix}$$

$$\hat{\mathbf{D}} = \begin{pmatrix} 0 & 9.0 & 4.7 & 10.2 & 6.8 \\ 9.0 & 0 & 4.7 & 3.0 & 3.9 \\ 4.7 & 4.7 & 0 & 7.9 & 5.8 \\ 10.2 & 3.0 & 7.9 & 0 & 4.7 \\ 6.8 & 3.9 & 5.8 & 4.7 & 0 \end{pmatrix}$$



The disparity matrix is obtained by conducting a monotone regression analysis on the configuration distance matrix. It meets the monotonicity criterion.

# From distance to disparity matrix

Distance matrix for data

$$\Delta = \begin{pmatrix} 0 & 9 & 4 & 10 & 7 \\ 9 & 0 & 3 & 1 & 2 \\ 4 & 3 & 0 & 8 & 6 \\ 10 & 1 & 8 & 0 & 5 \\ 7 & 2 & 6 & 5 & 0 \end{pmatrix}$$

Ordered distances of distance matrix

$$\delta_{24} < \delta_{25} < \delta_{23} < \delta_{13} < \delta_{45} < \delta_{35} < \delta_{15} < \delta_{34} < \delta_{12} < \delta_{14}$$

Ordered distances of disparity matrix

$$\hat{d}_{24} \leq \hat{d}_{25} \leq \hat{d}_{23} \leq \hat{d}_{13} \leq \hat{d}_{45} \leq \hat{d}_{35} \leq \hat{d}_{15} \leq \hat{d}_{34} \leq \hat{d}_{12} \leq \hat{d}_{14}$$

Disparity matrix

$$\hat{\mathbf{D}} = \begin{pmatrix} 0 & 9.0 & 4.7 & 10.2 & 6.8 \\ 9.0 & 0 & 4.7 & 3.0 & 3.9 \\ 4.7 & 4.7 & 0 & 7.9 & 5.8 \\ 10.2 & 3.0 & 7.9 & 0 & 4.7 \\ 6.8 & 3.9 & 5.8 & 4.7 & 0 \end{pmatrix}$$

# Goodness of fit for NMDS

STRESS 1 = $\sqrt{\frac{\sum_{i < j} (d_{i,j} - \widehat{d}_{i,j})^2}{\sum_{i < j} d_{i,j}^2}}$	Value of STRESS1	Goodness of configuration
	< 0.05	excellent
	< 0.10	good
	< 0.15	medium
	> 0.15	bad

## Implementation of NMDS in R

monoMDS() {vegan}      Basic function for NMDS

metaMDS() {vegan}      „Shotgun“ method

cmdscale() {stats}

cmds() {mclust}      (Metric) multidimensional scaling



# What does the “shotgun” method do?

`metaMDS()` {vegan}

1. Data transformation (Square-root and Wisconsin double transformation)
2. Calculation of distance matrix based on the selected similarity coefficient (defaults to Bray-Curtis)
3. Adjustment if no shared occurrences of species
4. Several random starts for initial configuration
5. Centring and rotation of ordination (highest dispersion on 1<sup>st</sup> axis)
6. Scaling (1 unit means halving of community similarity)
7. Calculation of species scores as weighted averages of sites

40

More information on ordination with metaMDS can be found in the corresponding help and in related documents of the vegan package:

<http://cran.r-project.org/web/packages/vegan/vignettes/intro-vegan.pdf>

# Limitations of NMDS

- Results dependent on initial configuration
- Loss of information due to ordered rank ordination
  - Information on absolute distances lost
  - No partitioning of variance
- Interpretation difficult if more than 2 or 3 dimensions required (i.e. to yield low STRESS1 value)
- Significant fit of environmental variables to ordered distances more difficult to interpret than for unconstrained variance-based methods