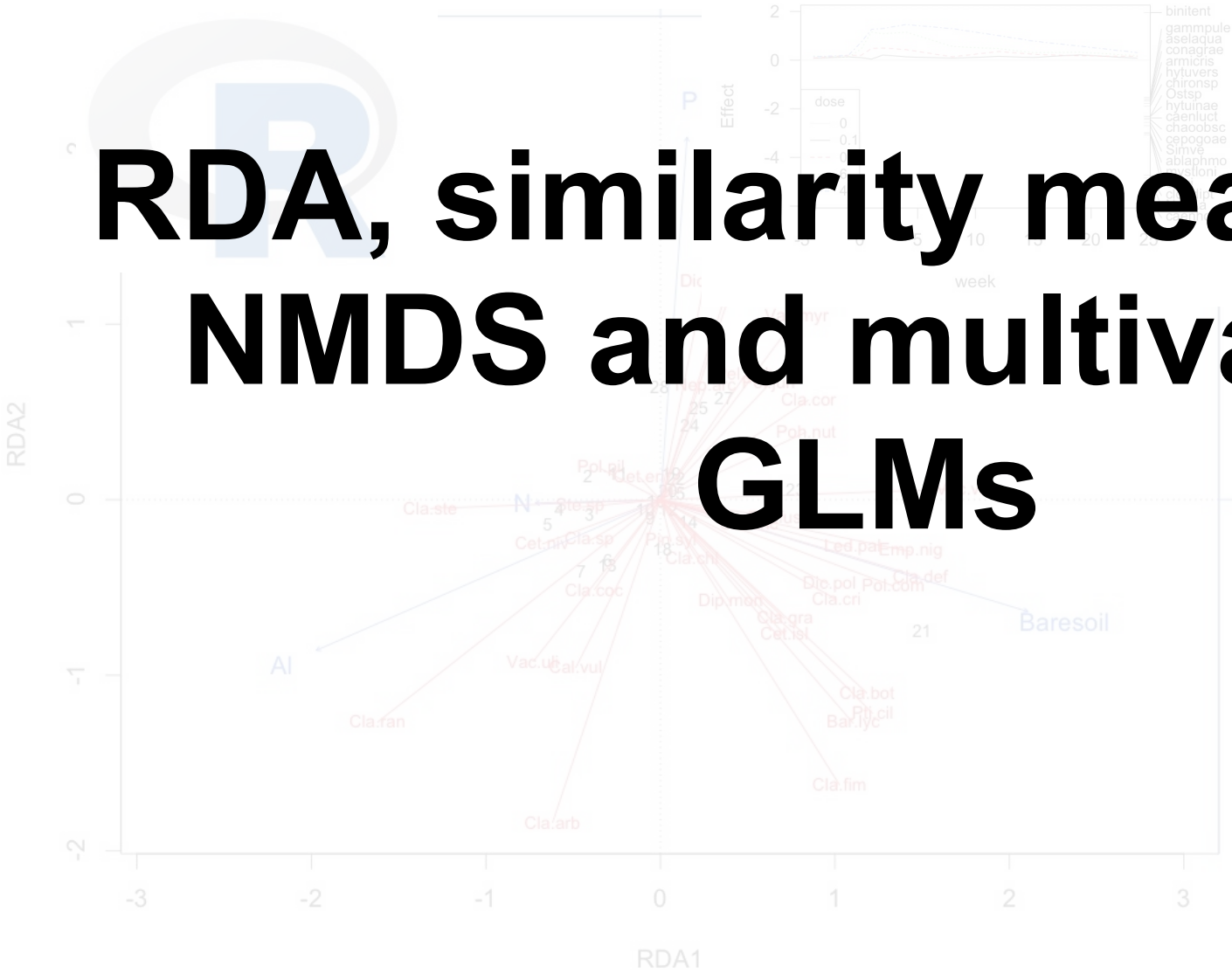
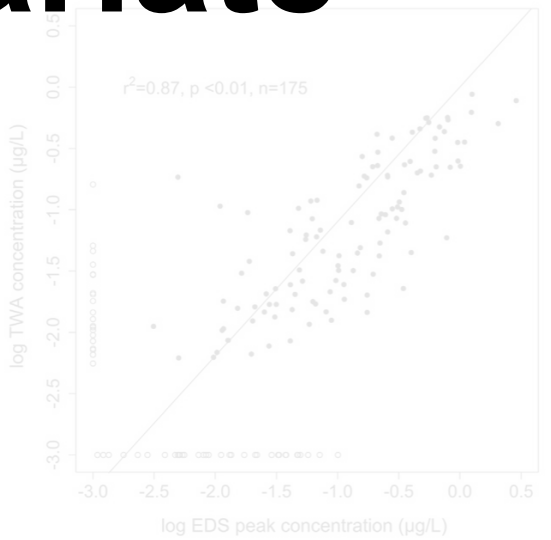
Tools for complex data analysis

University of Koblenz-Landau 2020/21





RDA, similarity measures, NMDS and multivariate GLMs

Ralf B. Schäfer

Learning targets

- Understanding the basics of RDA and extensions.
- Knowledge on the calculation of commonly used association measures.
- Understanding the mathematical background and how to conduct a NMDS.
- Knowledge on model-based ordination and multivariate GLMs

Learning targets and study questions

- Understanding the basics of RDA and extensions.
  - How many constrained axes has an RDA and how are they related to the descriptors?
  - What can be interpreted in a RDA triplot?
  - How is PRC related to RDA and for which research contexts is it useful?
- Knowledge on the calculation of commonly used association measures.
  - Which association is measured with similarity measures?
  - For which data are the Bray-Curtis and the Jaccard coefficient suitable?
  - How can similarity measures be visualised?

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 types of matrices are computed during NMDS?
  - Outline the main steps when computing a NMDS.
  - Discuss limitations of NMDS.
- Knowledge on model-based ordination and multivariate GLMs
  - Discuss the limitations of dissimilarity-based and algorithm-based methods. How do model-based methods overcome these limitations?
  - Explain the basics of multivariate GLMs.

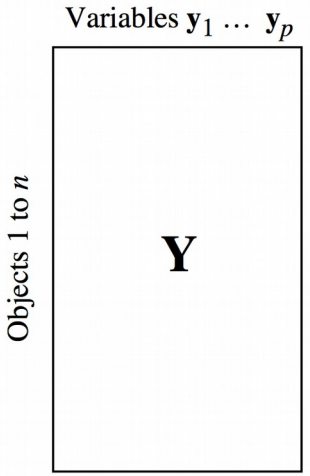
RDA, similarity measures, NMDS and multivariate GLMs

Contents

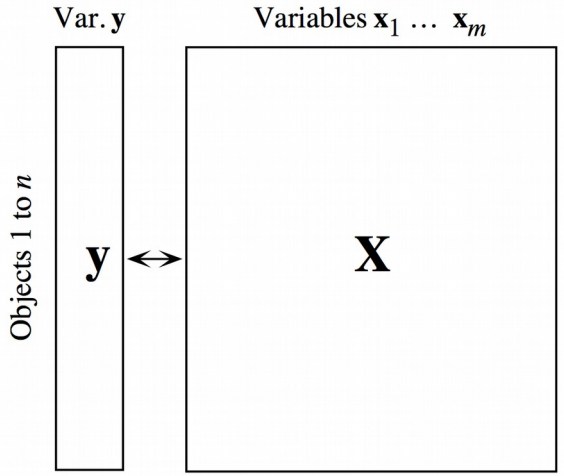
1. **Constrained ordination and RDA**
2. RDA assumptions and PRC
3. Similarity measures
4. Nonmetric multidimensional scaling (NMDS)
5. Model-based ordination, Multivariate GLMs

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

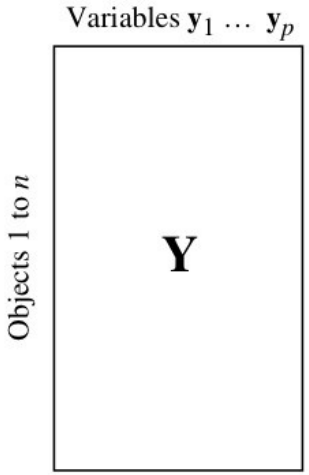


Model: $\hat{y} = b_0 + b_1 x_1 + \ldots + b_m x_m$

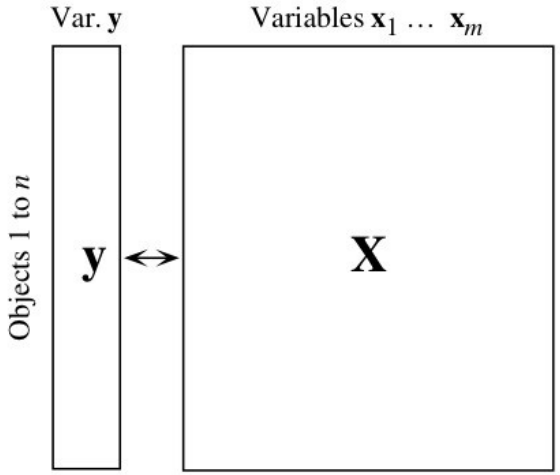
Legendre & Legendre 2012: 627

Constrained ordination methods

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

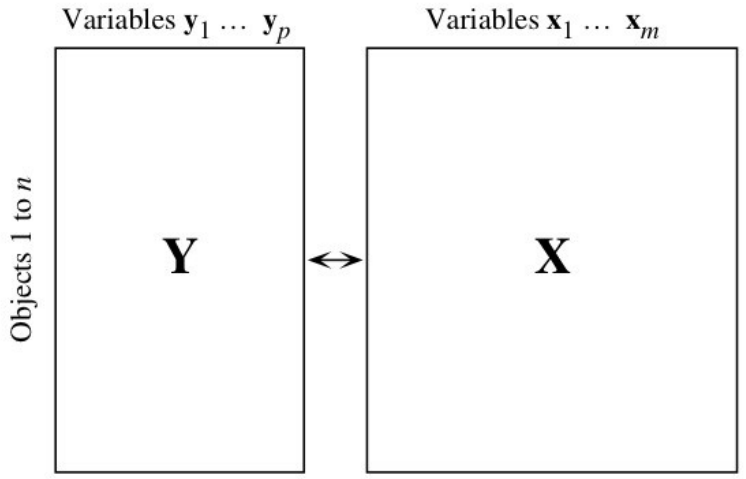


(b) Ordination of **y** (single axis) under
constraint of **X**: multiple regression



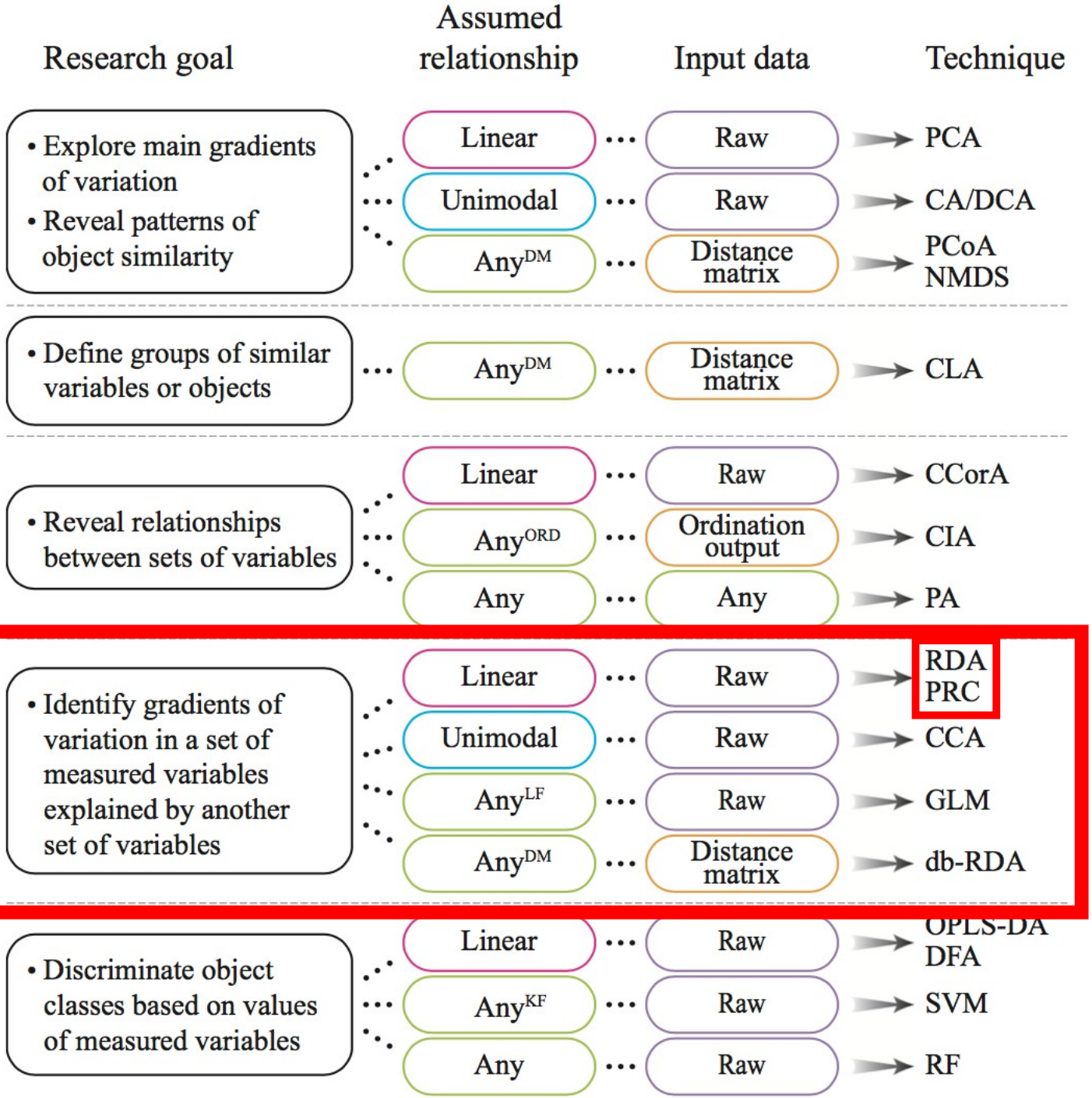
Model: $\hat{y} = b_0 + b_1 x_1 + \ldots + b_m x_m$

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



Legendre & Legendre 2012: 627

Constrained ordination methods

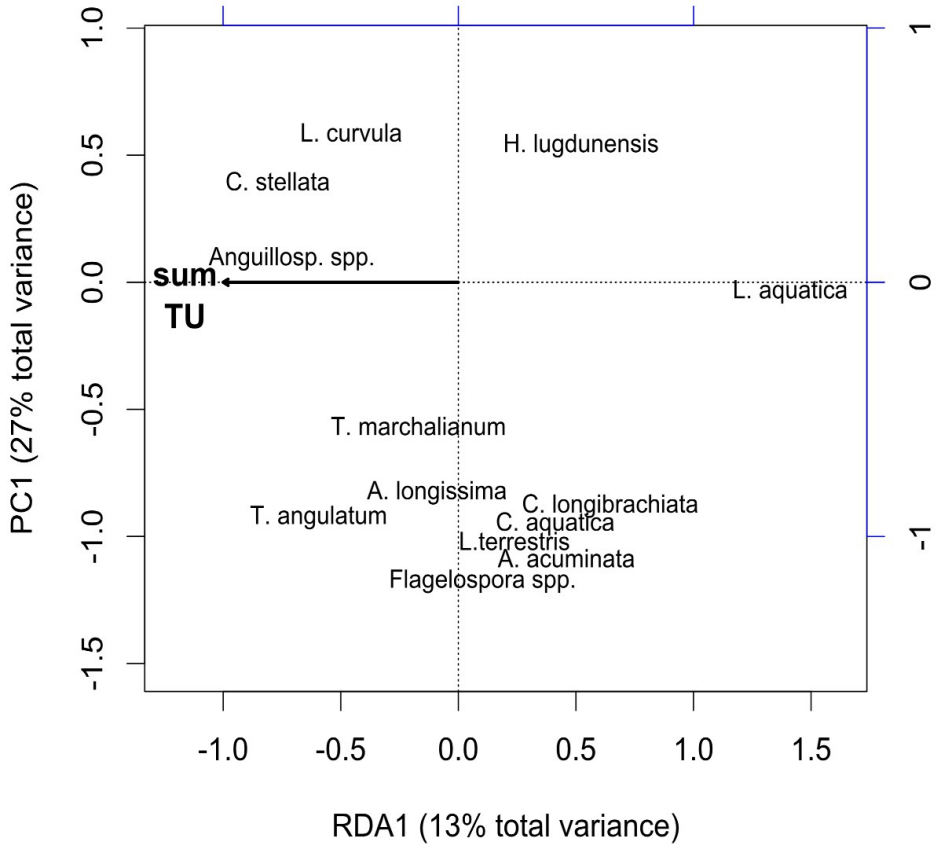


Paliy & Shankar 2016 *Mol Ecol* 25: 1032

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?



Fernandez et al. 2015 *Sci. Tot. Environ.* 533: 40

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

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

$$\hat{y} = \boldsymbol{X} b$$

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

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

Reformulation for multivariate multiple regression with several y:

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

Mathematical background of RDA

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

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

Usually, this is not known and the sample variance-covariance matrix $\boldsymbol{S}$ (also called Dispersion matrix) is estimated from the observations:

$$\boldsymbol{S}_{\hat{\boldsymbol{Y}}^T\hat{\boldsymbol{Y}}} = \frac{1}{n-1}\hat{\boldsymbol{Y}}^T\hat{\boldsymbol{Y}}$$

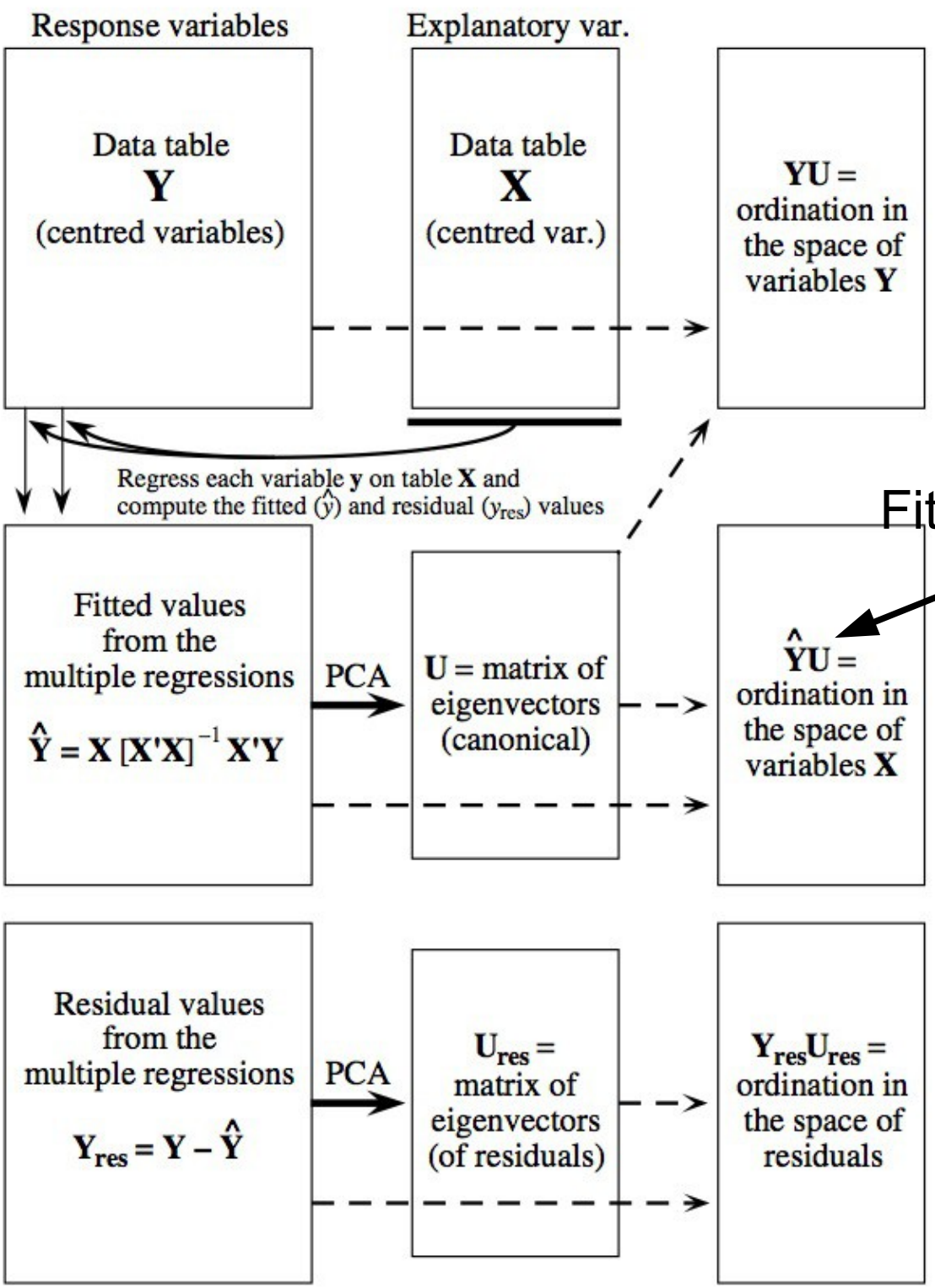
and used in a PCA:

$$\boldsymbol{S}_{\hat{\boldsymbol{Y}}^T\hat{\boldsymbol{Y}}}\,a = \lambda\, a$$

Eigenvector

Eigenvalue problem

➡ Eigenvectors linear combinations of predictors



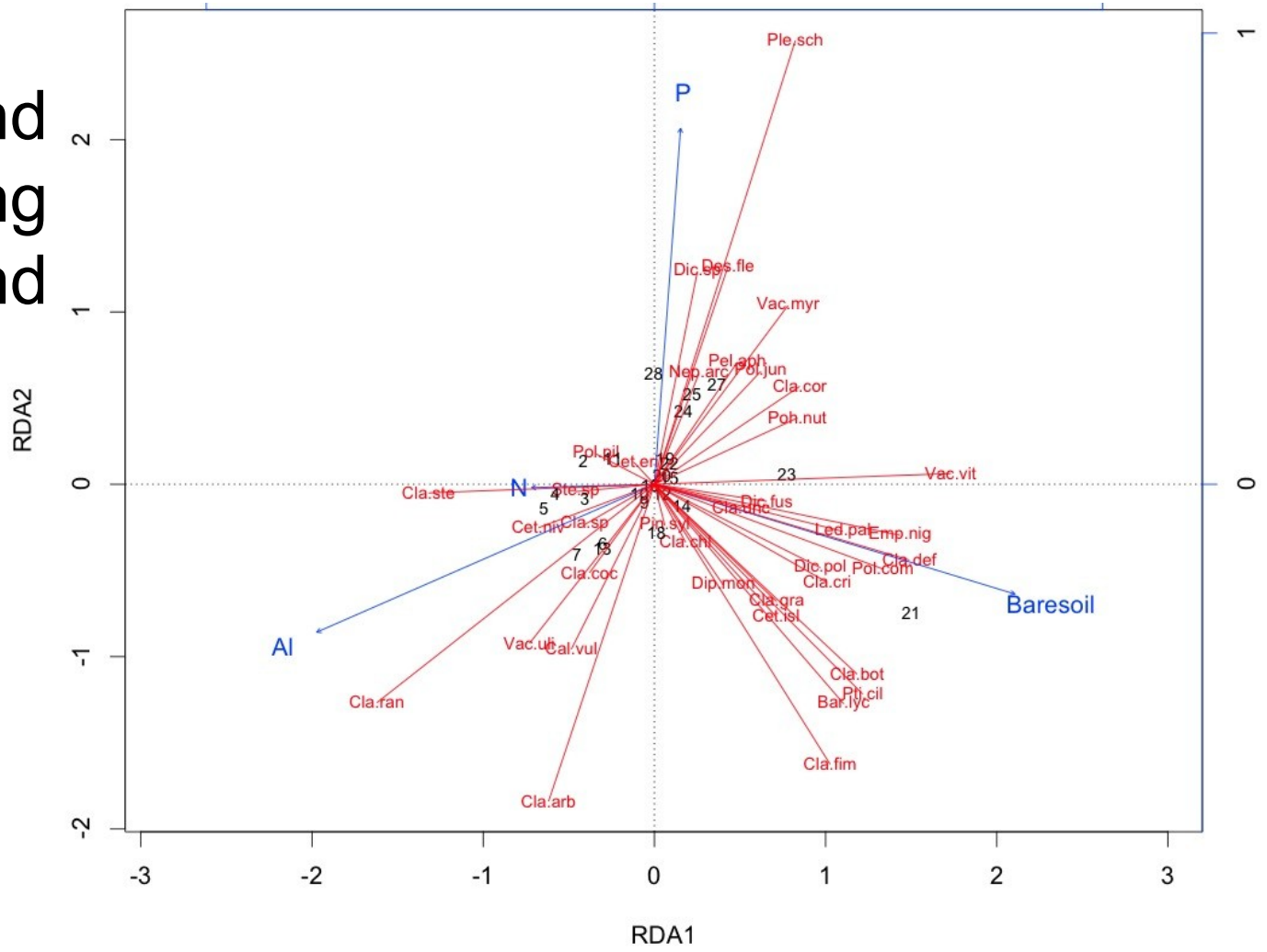
Fitted site scores

Legendre & Legendre 2012:631

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, NMDS and multivariate GLMs

Contents

1. Constrained ordination and RDA
2. **RDA assumptions and PRC**
3. Similarity measures
4. Nonmetric multidimensional scaling (NMDS)
5. Model-based ordination, Multivariate GLMs

RDA model and variable importance

How many RDA axes are required?

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

Which explanatory variables should be included in the best-fit RDA if the research goal is explanation and how important are they?

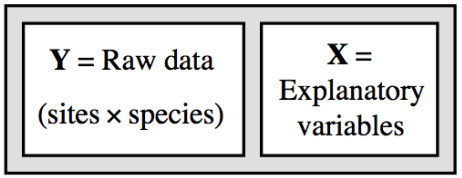
- Manual and automatic model-building with adjusted R^2 as goodness of fit criterion (as for multiple linear regression) and permutation-based p-values
- Variance partitioning between different models to determine explained variance of individual explanatory variables

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 (objects) >> p (predictors/explanatory variables) to reliably infer p importance
- RDA can be employed for multivariate ANOVA (see Borcard et al. 2018: 238 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 extensions

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



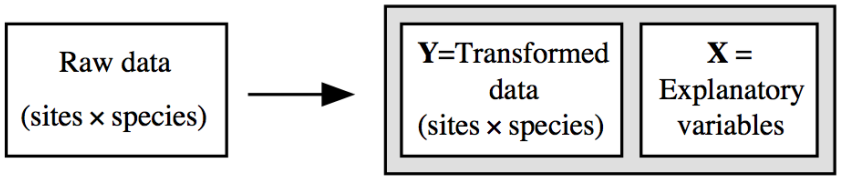
Short gradients: CCA or RDA

Long gradients: CCA

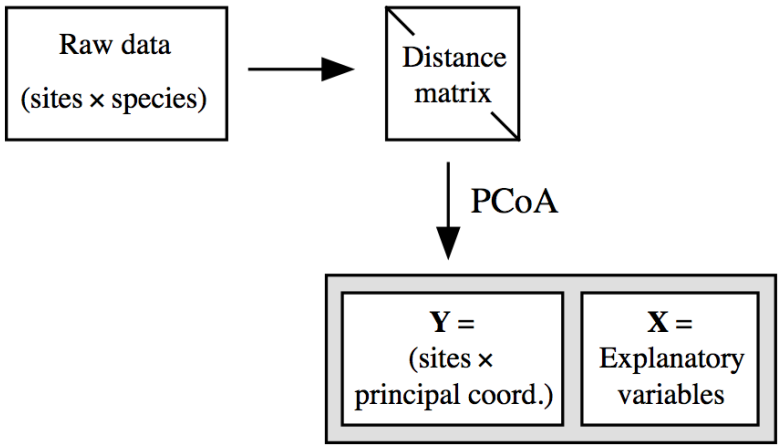
How to assess gradient length?

- test for higher order terms (Borcard et al. 2018: 244ff)
- Axis length in DCA

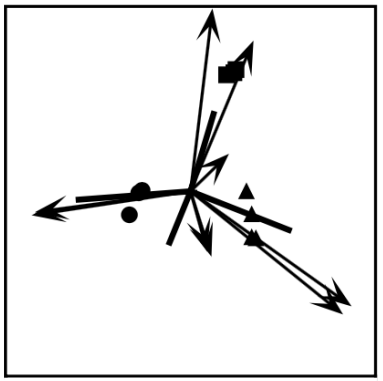
(b) Transformation-based RDA (tb-RDA) approach:
preserves a distance obtained by data transformation



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



Canonical ordination triplot



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

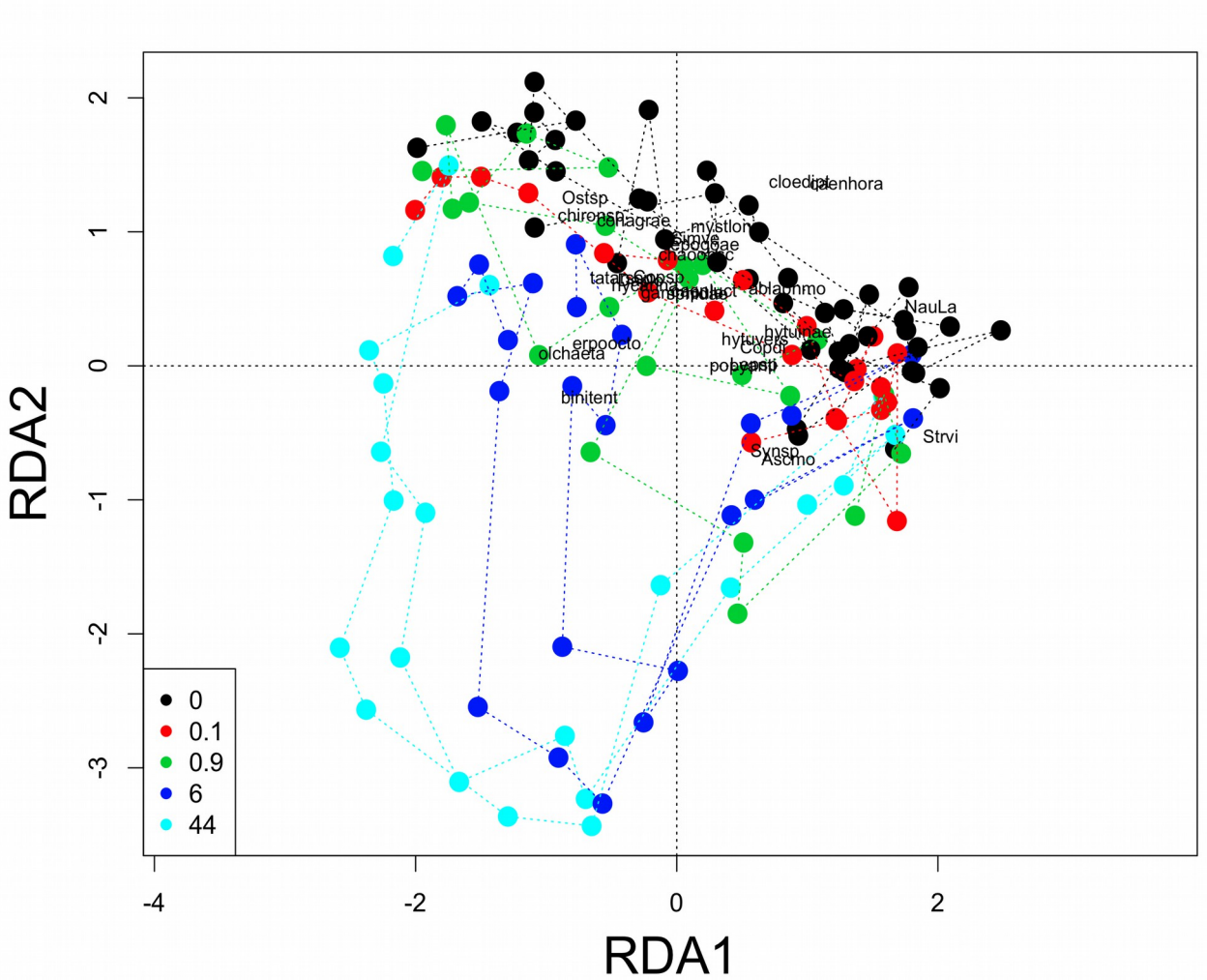
d) Alternative approach:
Model-based ordination (e.g. multivariate GLMs, CAO)

Legendre & Legendre 2012: 648

RDA extension: Principal Response Curve (PRC)

Example: Before-After-Control-Impact (BACI) study with communities. Treatment of aquatic mesocosms containing invertebrates with insecticide chlorpyrifos.

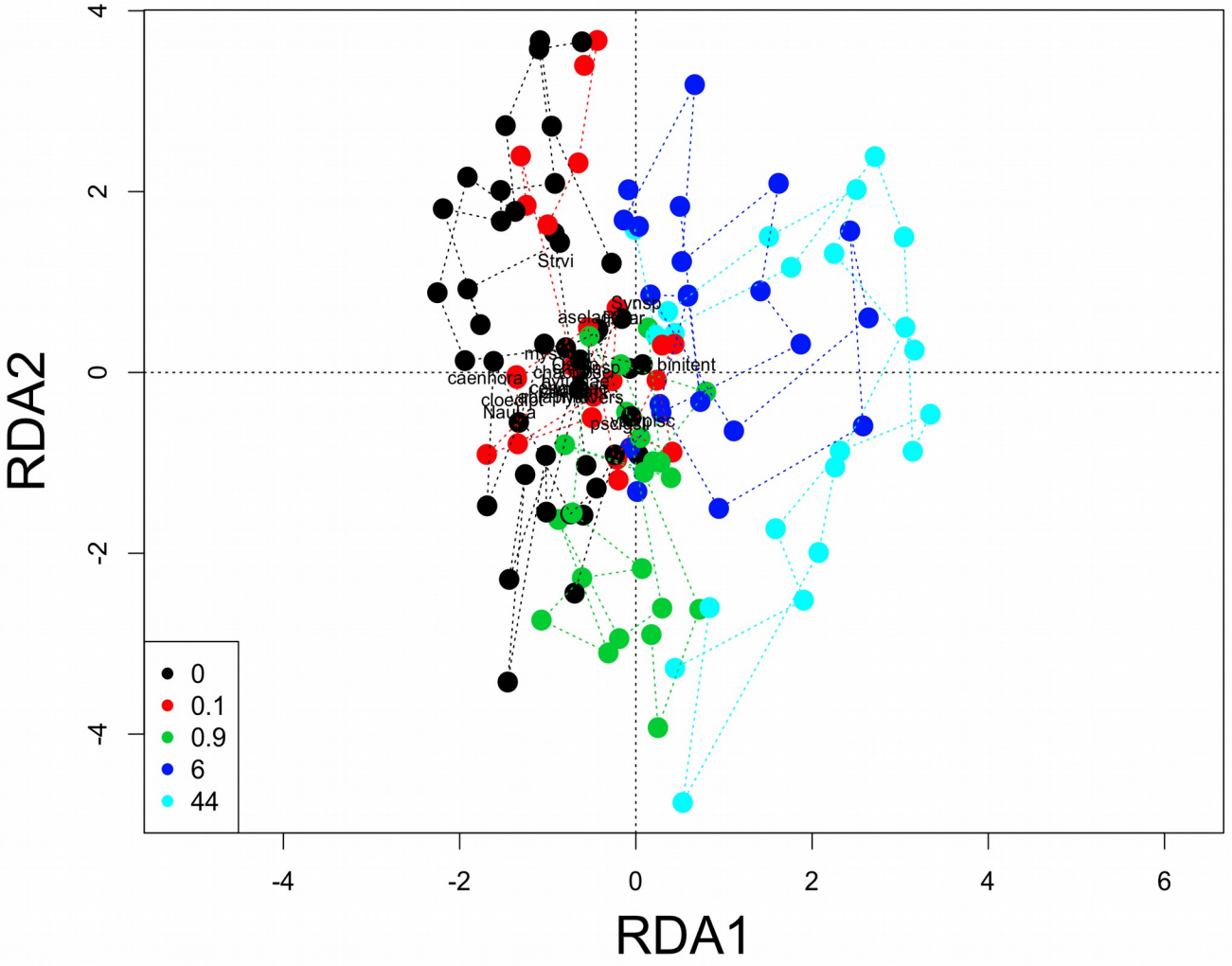
- RDA model for time, treatment and their interaction
- Clear time and treatment effect but figure cluttered
- Pure time effect often not relevant → Remove from model



RDA extension: Principal Response Curve (PRC)

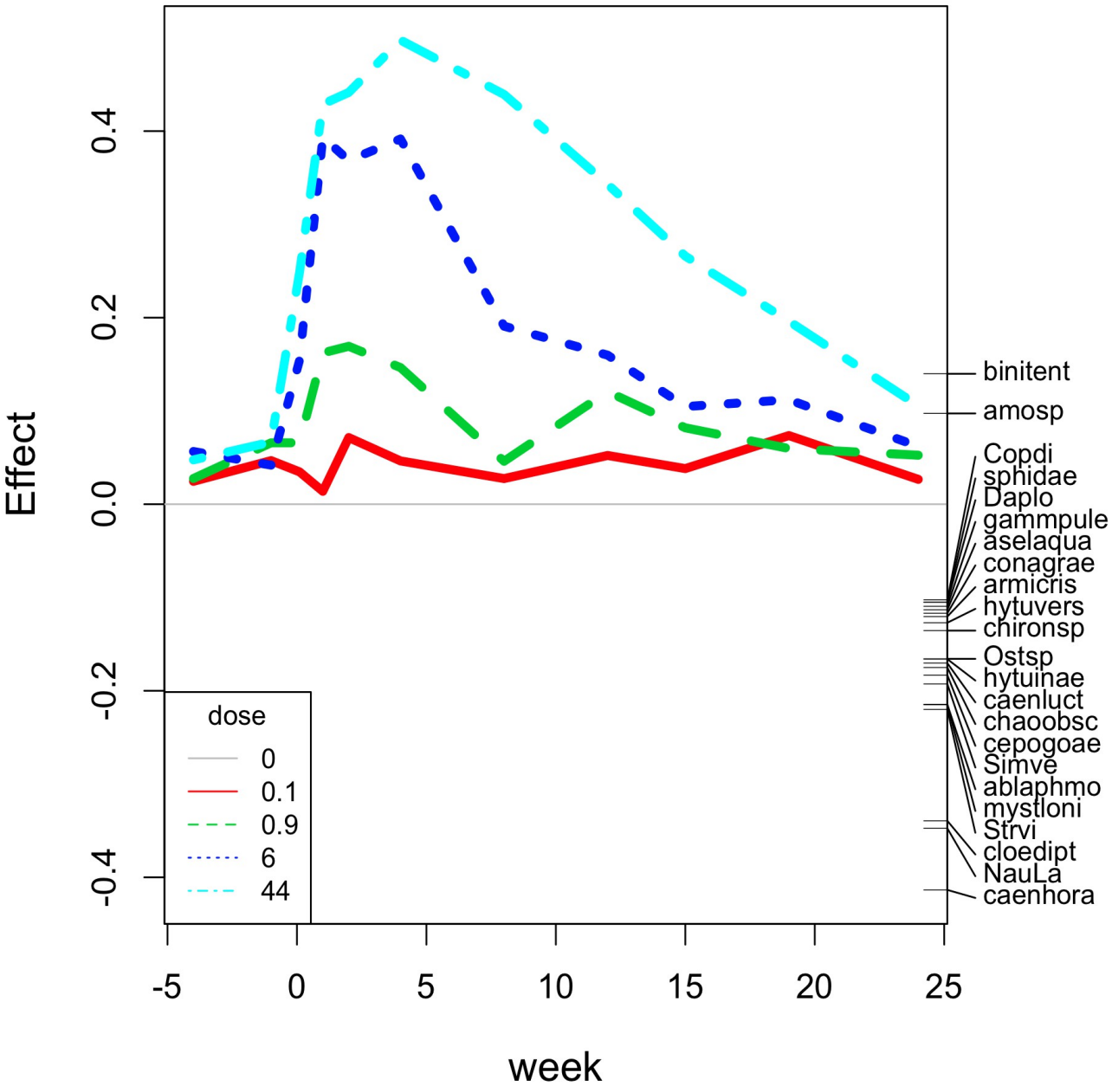
Partial RDA: Explanatory variables held constant to remove their effect (Borcard et al. 2018: 221-225)

- Time effect removed ("partialled out") with partial RDA
- First axis: Treatment and interaction effect
- Better separation between treatments, still cluttered
  → Focus on 1st axis

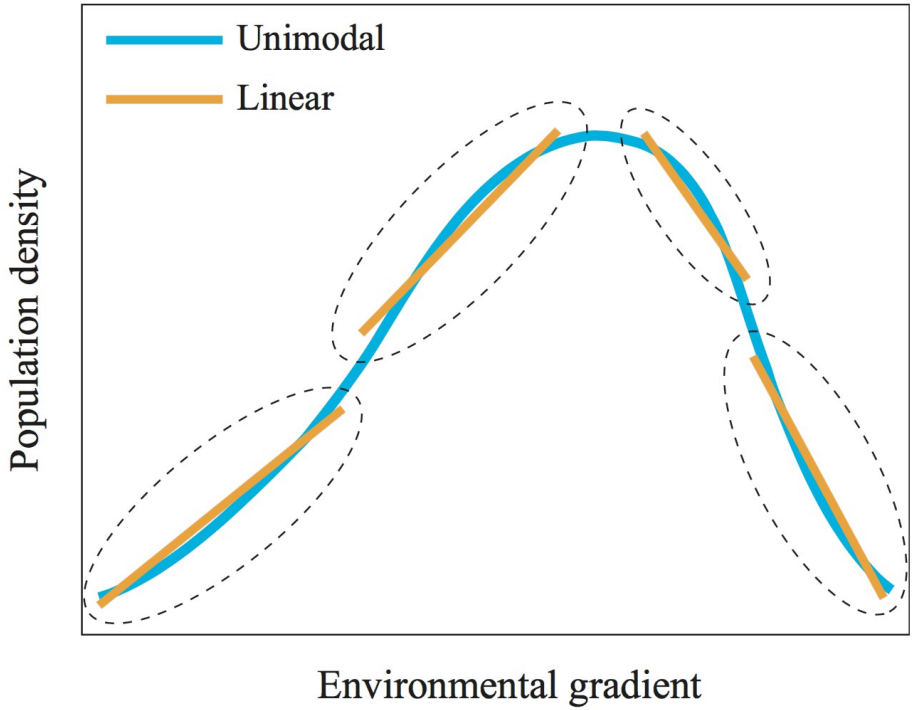


Principal Response Curve (PRC)

- 1st axis of partial RDA
- Treatments plotted relative to control by subtracting site scores
- Y axis: Treatment effect (difference in composition)
- X axis: Time
- Species scores: Species responsible for pattern
- Recovery if treatments approach control



Alternative to RDA for unimodal responses



Canonical Correspondence Analysis (CCA)

- Similar to RDA, but assumes unimodal distribution (χ^2-distance) of species along environmental gradient → widely used for ecological data
- Extension of (unconstrained) correspondence analysis
- Similar modelling framework as for RDA

Paliy & Shankar 2016 *Mol Ecol*:1032

RDA, similarity measures, NMDS and multivariate GLMs

Contents

1. Constrained ordination and RDA
2. RDA assumptions and PRC
3. **Similarity measures**
4. Nonmetric multidimensional scaling (NMDS)
5. Model-based ordination, Multivariate GLMs

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

Similarity measures for occurrence data

Simple matching coefficient

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

$$S_{\text{Match}} = \frac{a+d}{a+b+c+d}$$

Exercise: Calculate S_{Match} for the data below with and without the 1. and 4. species. How do these species influence S_{Match}?

Site				
1	0	400	0	0
2	0	0	10	0

Similarity measures for occurrence data

Site				
1	0	400	0	0
2	0	0	10	0

$$S_{\text{Match}} = \frac{a+d}{a+b+c+d}$$

Calculation with all species:

$a = 0,\ b = 1, c = 1, d = 2 \rightarrow S_{\text{Match}} = 2/4 = 0.5$

Calculation without species 1 and 4:

$a = 0,\ b = 1, c = 1, d = 0 \rightarrow S_{\text{Match}} = 0/2 = 0$

Joint absence of species influences similarity between sites

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

→ **Double-zero problem**

Widely used similarity measures

Jaccard coefficient S_{Jacc} (=Jaccard similarity index)

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

$$S_{\text{Jacc}} = \frac{a}{a+b+c}$$

- occurrence data
- ignores joint absences (d)
- Range: 0 (no similarity) to 1 (identity)

Bray-Curtis coefficient S_{BC}

- abundance data
- Range: 0 to 1
- Often prior data transformation to reduce weight of dominant taxa

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

$x_{i,k}$ and $x_{j,k}$ is the abundance of taxon k in site i and j.

Example: Transformation and S_{BC}

Site				
i = 1	0	400	5	0
j = 2	0	0	10	0
Min	0	0	5	0
Sum	0	400	15	0

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

Calculation:

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

Calculation for square-root transformed data:

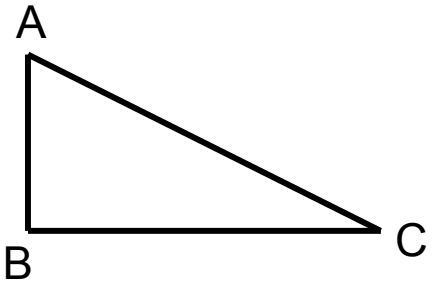
2*(0+0+$5^{0.5}$+0)/($400^{0.5}$+$5^{0.5}$+$10^{0.5}$) $\rightarrow S_{BC}$ = 0.18

Calculation for double square-root transformed data:

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

Distance measure

Association (e.g. dissimilarity) measure meeting triangle inequality criterion:



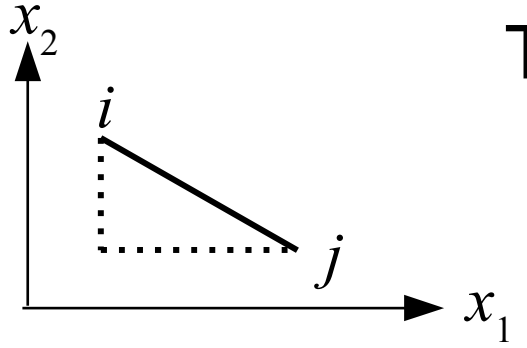
d(A,B) + d(B,C) ≥ d(A,C), where d is distance function

Sum of any two sides of triangle always ≥ third side

Important for geometrical representation (e.g. ordination)

<u>Euclidean distance:</u> frequently used distance measure (e.g. PCA, RDA), not suitable for ecological data (→ species abundance paradox)

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



Two dimensions:

$$D_{\mathrm{Eucl}}(i,j) = \sqrt{(x_{i,1}-x_{j,1})^2+(x_{i,2}-x_{j,2})^2}$$

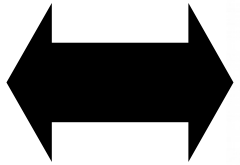
Visualization of association measures

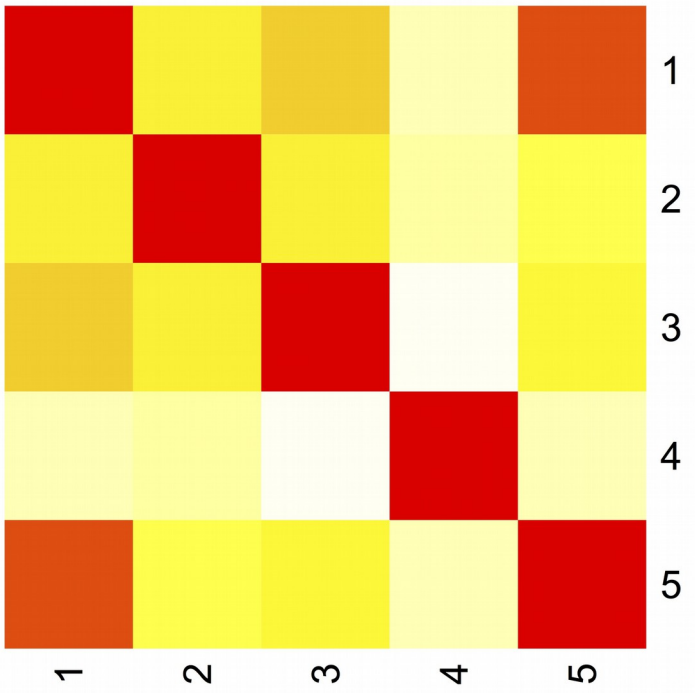
Heatmap

- Associations converted to colours
- Relationship easier to grasp

Matrix of distances between sites

	1	2	3	4	5
1	0	0.69	0.6	0.92	0.22
2	0.69	0	0.7	0.89	0.8
3	0.6	0.7	0	0.98	0.72
4	0.92	0.89	0.98	0	0.92
5	0.22	0.8	0.72	0.92	0



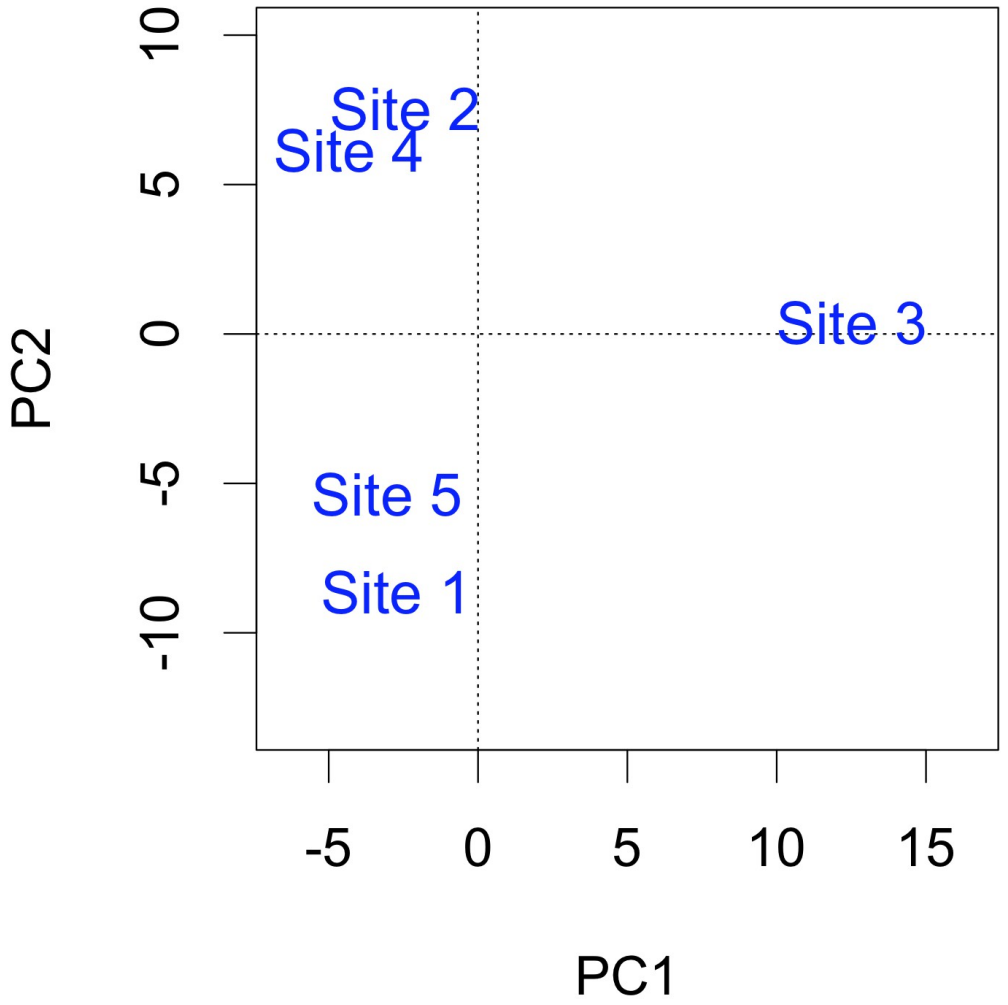


Diagonal entries: Distance of site to itself (= 0)

Visualization of association measures

Ordination

Measures that meet triangle inequality criterion allow for clear geometrical interpretation of ordination



How to select an association 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: *simple matching* (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**

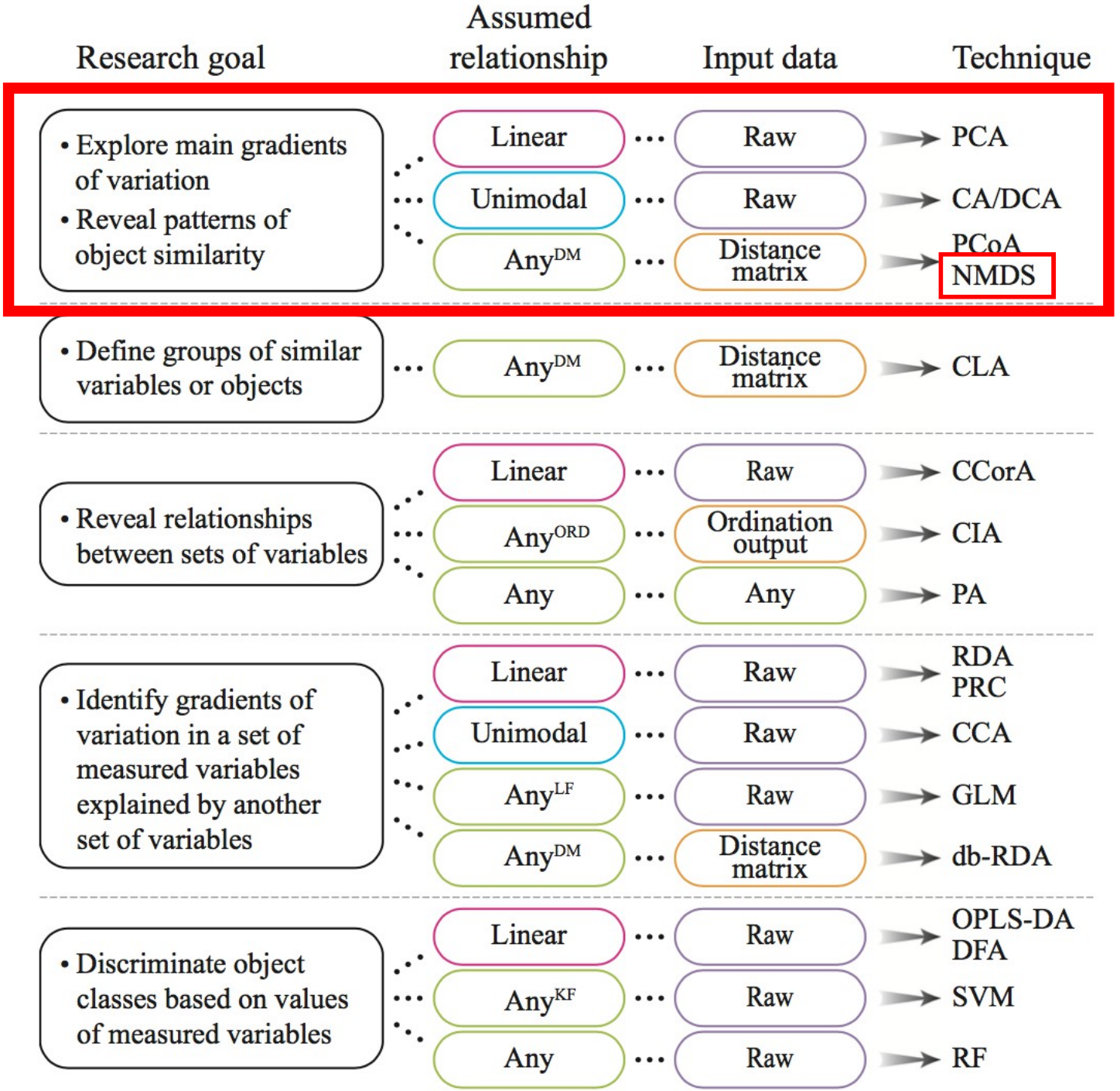
5) Differences enhanced by squaring: *Euclidean distance* (D_1) and *average distance* (D_2)

RDA, similarity measures, NMDS and multivariate GLMs

Contents

1. Constrained ordination and RDA
2. RDA assumptions and PRC
3. Similarity measures
4. **Nonmetric multidimensional scaling (NMDS)**
5. Model-based ordination, Multivariate GLMs

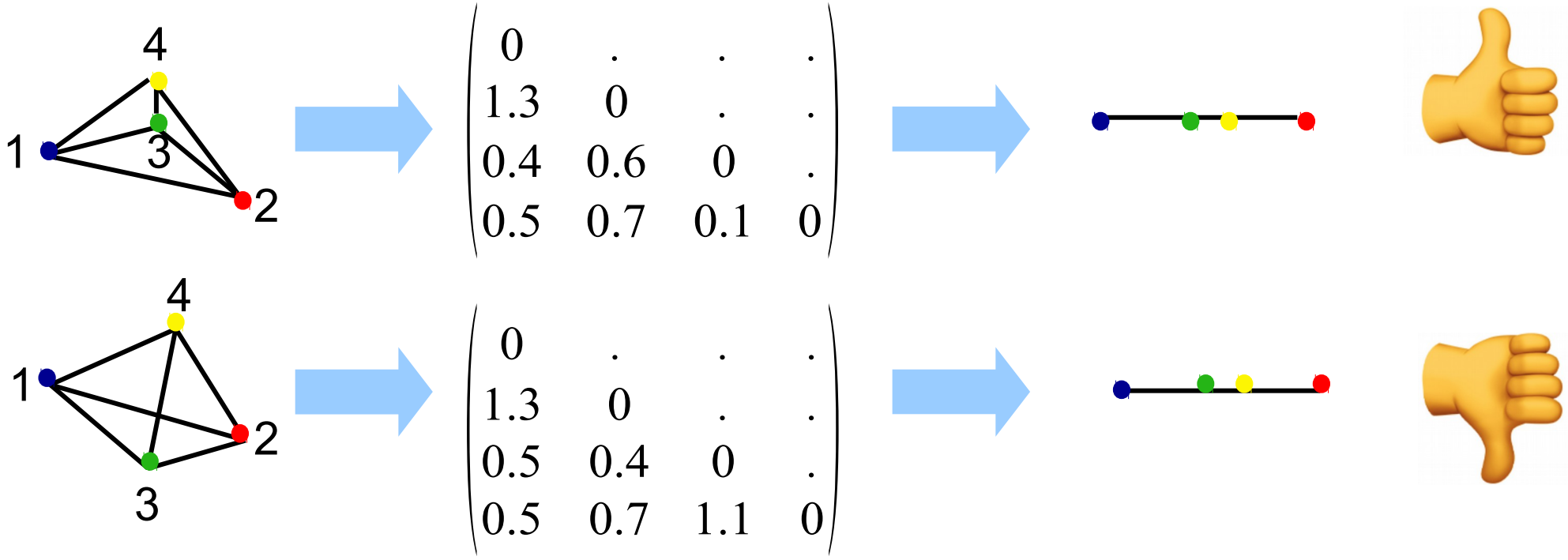
Unconstrained ordination with NMDS



Paliy & Shankar 2016 *Mol Ecol*:1032

Challenge of lower-dimensional preservation of metric distances

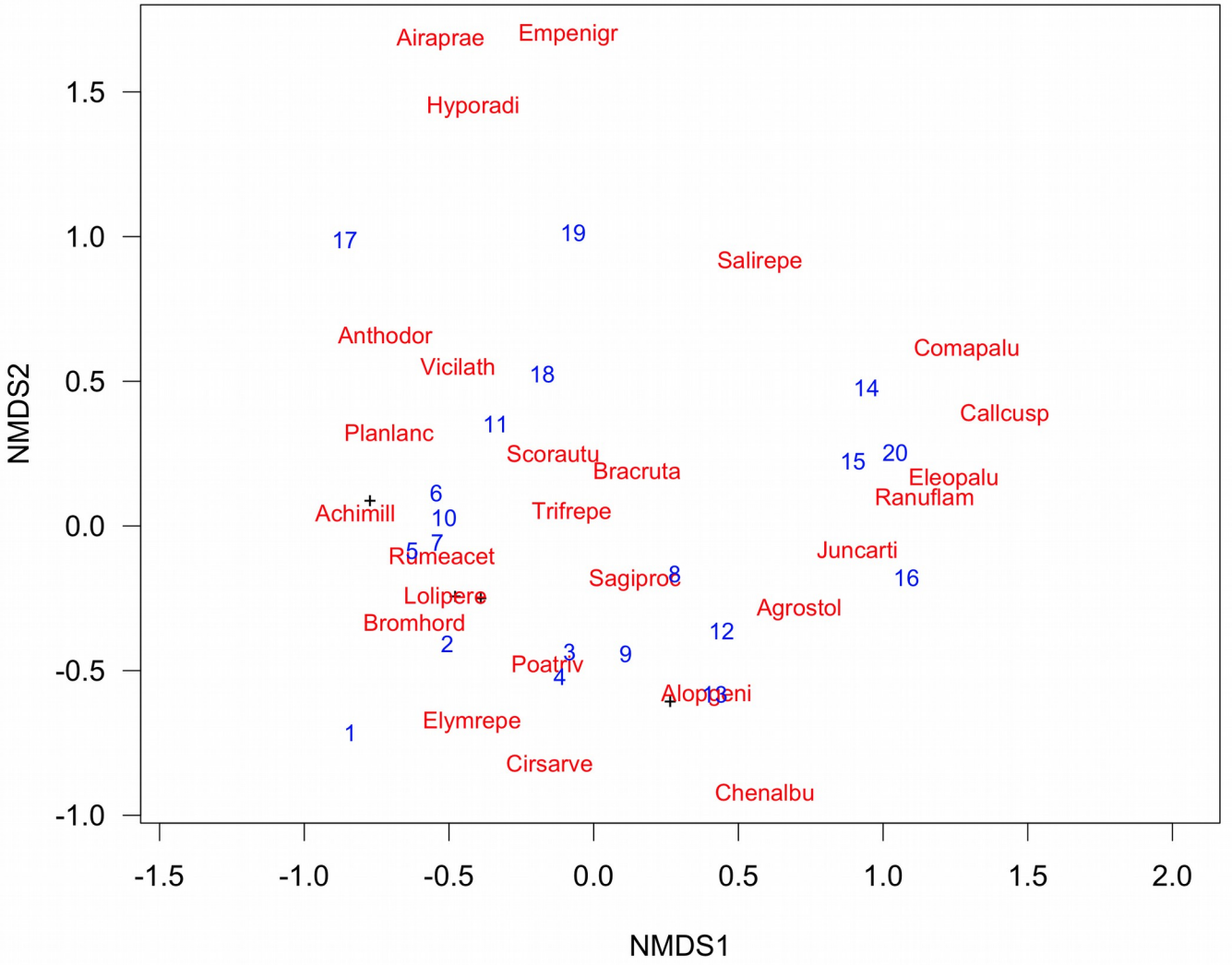
Example for one-dimensional representation of two dimensions:



Idea: Relax condition for lower dimensional representation by preserving rank instead of absolute distances

Nonmetric multidimensional scaling

- Unconstrained ordination based on ordered ranks of pairwise distances or dissimilarities (→ nonmetric)
- Can be used with many distance/ dissimilarity measures
  → Suitable for ecological data
- Not based on eigenvalues, no partitioning of variance
- Very robust and flexible



Steps of NMDS algorithm

1. Determine distance matrix Δ for raw data
2. Set number of dimensions k
3. Set initial configuration
4. Determine distance matrix D for configuration
5. Monotone regression and Pool Adjacent Violators (PAV) algorithm $\rightarrow$ Disparity matrix $\hat{\mathrm{D}}$ and Goodness of fit measure $\mathrm{STRESS1}$
6. Start with new random configuration and go to 4. (if fit does not improve on many iterations $\rightarrow$ 7.)
7. Final configuration

From distance matrix for raw data...

1. Determine distance matrix Δ for raw data

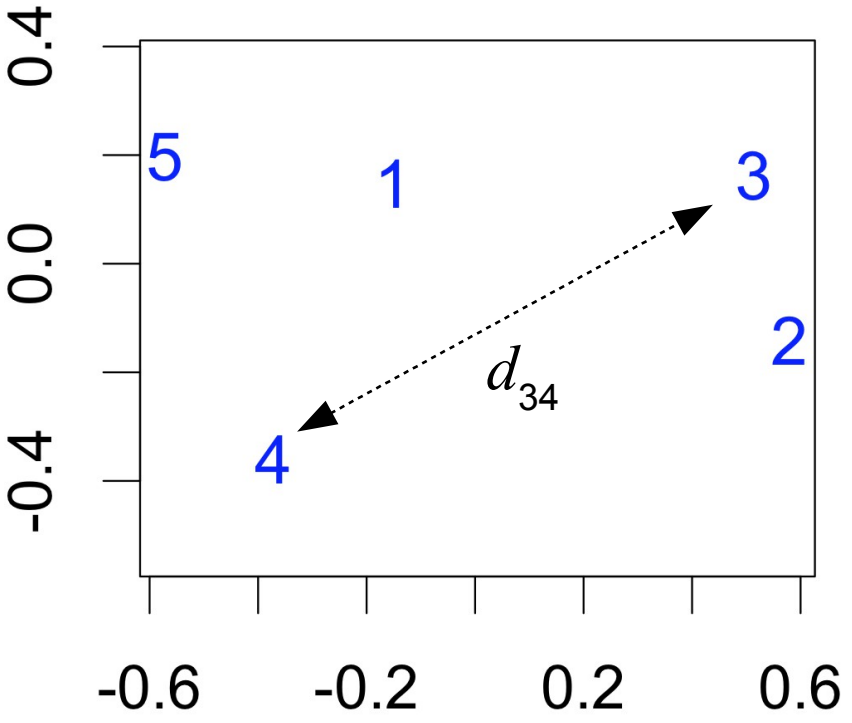
$$\Delta = \begin{pmatrix} 0 & 9 & 4 & 10 & 7 \\ 9 & 0 & 3 & 1 & 2 \\ 4 & 3 & 0 & 8 & 6 \\ 10 & 1 & 8 & 0 & 5 \\ 7 & 2 & 6 & 5 & 0 \end{pmatrix}$$

Ordered distances of distance matrix

$$\delta_{24} < \delta_{25} < \delta_{23} < \delta_{13} < \delta_{45} < \delta_{35} < \ldots$$

2. Set number of dimensions: $k = 2$

3. Set initial configuration



… over distance matrix for configuration...

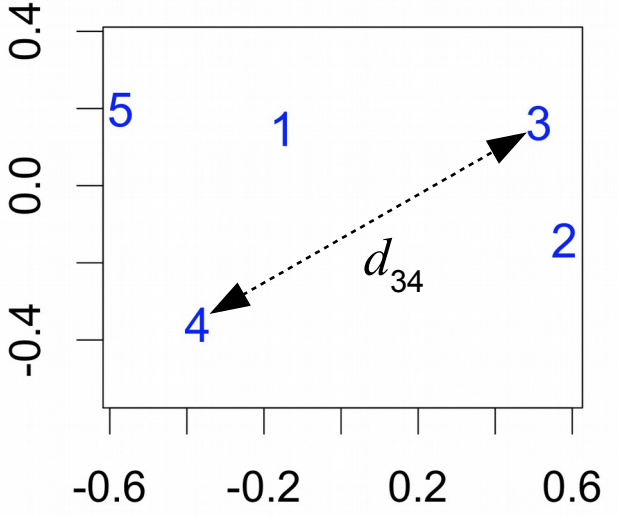
1. Determine distance matrix Δ for raw data

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

Ordered distances of distance matrix

$$\delta_{24} < \delta_{25} < \delta_{23} < \delta_{13} < \delta_{45} < \delta_{35} < \ldots$$

4. Determine distance matrix D for configuration



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

$$d_{24} < d_{45} < d_{25} < d_{13} < d_{23} < d_{35} < \ldots$$

Order of distances of D not matching with Δ

… to disparity matrix

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

$$d_{24} < d_{45} < d_{25} < d_{13} < d_{23} < d_{35} < \dots$$

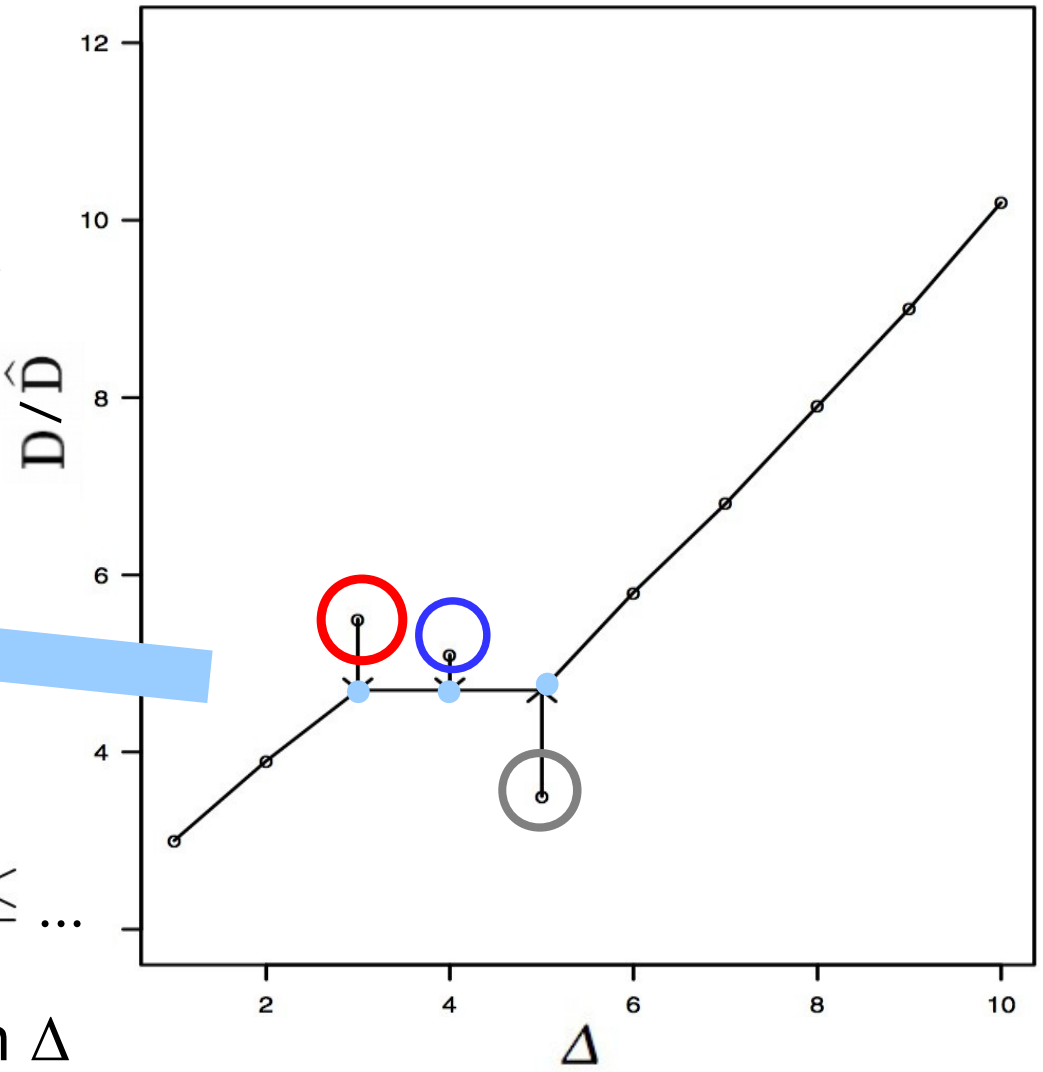
$$\widehat{\mathbf{D}} = \begin{pmatrix} 0 & 9.0 & 4.7 & 10.2 & 6.8 \\ 9.0 & 0 & 4.7 & 3.0 & 3.9 \\ 4.7 & 4.7 & 0 & 7.9 & 5.8 \\ 10.2 & 3.0 & 7.9 & 0 & 4.7 \\ 6.8 & 3.9 & 5.8 & 4.7 & 0 \end{pmatrix}$$

$$\hat{d}_{24} \leq \hat{d}_{25} \leq \hat{d}_{23} \leq \hat{d}_{13} \leq \hat{d}_{45} \leq \hat{d}_{35} \leq \dots$$

Order of distances of $\widehat{\mathbf{D}}$ matching with Δ

$$\delta_{24} < \delta_{25} < \delta_{23} < \delta_{13} < \delta_{45} < \delta_{35} < \dots$$

5. Monotone regression and PAV algorithm



Handl 2010: 174

Goodness of fit and number of dimensions

- Goodness of fit metric STRESS1: Difference between original distance and distance of final configuration (in $\hat{D}$)

$$\text{STRESS1} = \sqrt{\frac{\sum_{i<j} (d_{i,j} - \hat{d}_{i,j})^2}{\sum_{i<j} d_{i,j}^2}}$$

- Rules of thumb for interpretation (Clarke 1993)

Value of STRESS1	Goodness of configuration
< 0.05	excellent
< 0.10	good
< 0.2	medium
> 0.2	bad

- Problem: STRESS1 dependent on factors such as data type, sample size, dimensionality.
  → Alternative: Permutation-based assessing of hypotheses (Dexter et al. 2018)

Which number of dimensions to set?

- Main purpose of NMDS is visualisation: 2-3 dimensions
- Use thresholds related to STRESS1 (but see Dexter et al. 2018)

Limitations of NMDS

- Results dependent on initial and random configurations
- 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 (e.g. to yield lower STRESS1 value)
- Fit of environmental variables more difficult to interpret than for metric (unconstrained) methods

RDA, similarity measures, NMDS and multivariate GLMs

Contents

1. Constrained ordination and RDA
2. RDA assumptions and PRC
3. Similarity measures
4. Nonmetric multidimensional scaling (NMDS)
5. **Model-based ordination, Multivariate GLMs**

Dissimilarity- and algorithm-based ordination

Y = Response (sites x species) transformed or non-transformed

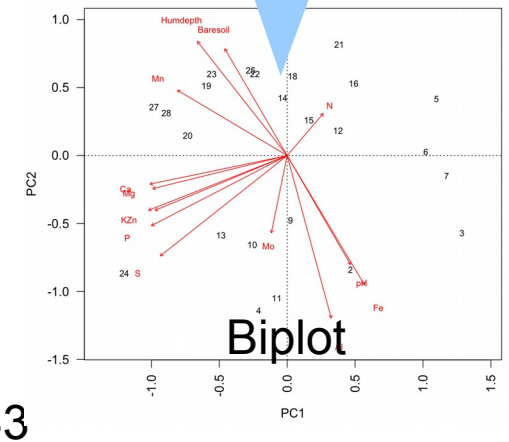
X = Predictors (sites x environmental vars.)

Unconstrained ordination (e.g. PCA, NMDS)

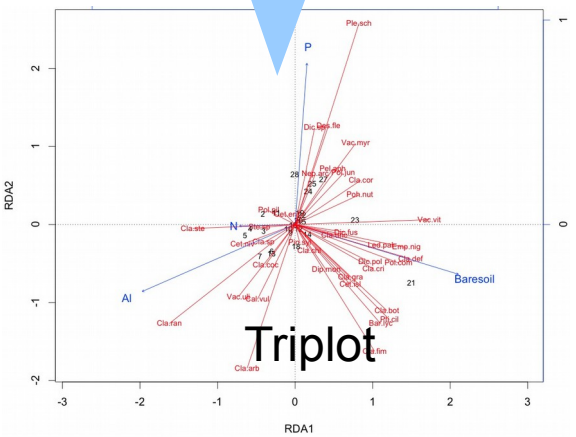
Constrained ordination (e.g. db-RDA, CCA)

$$A\,x = \lambda\,x$$ Matrix decomposition

Algorithms on dissimilarity matrix



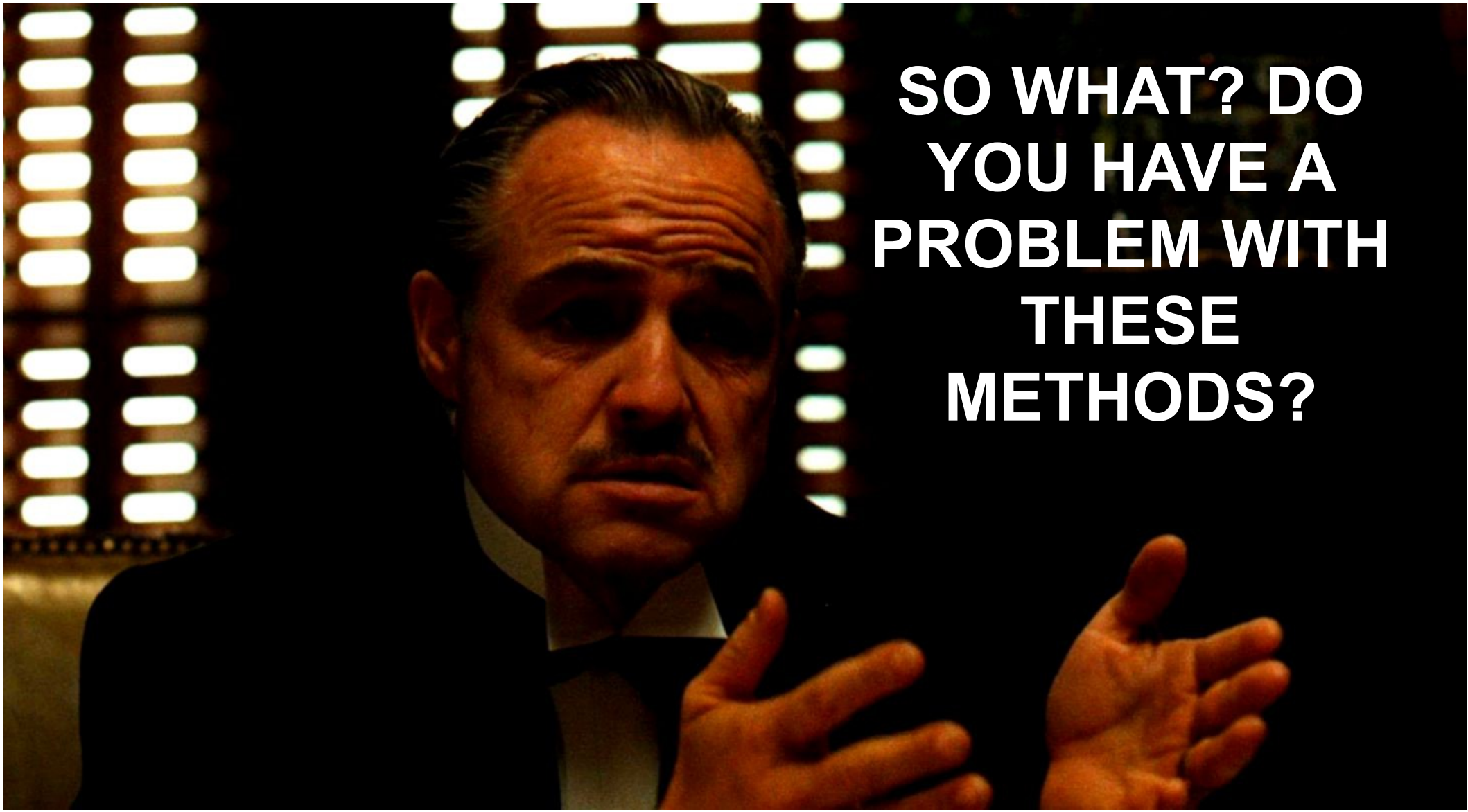
Biplot



Triplot

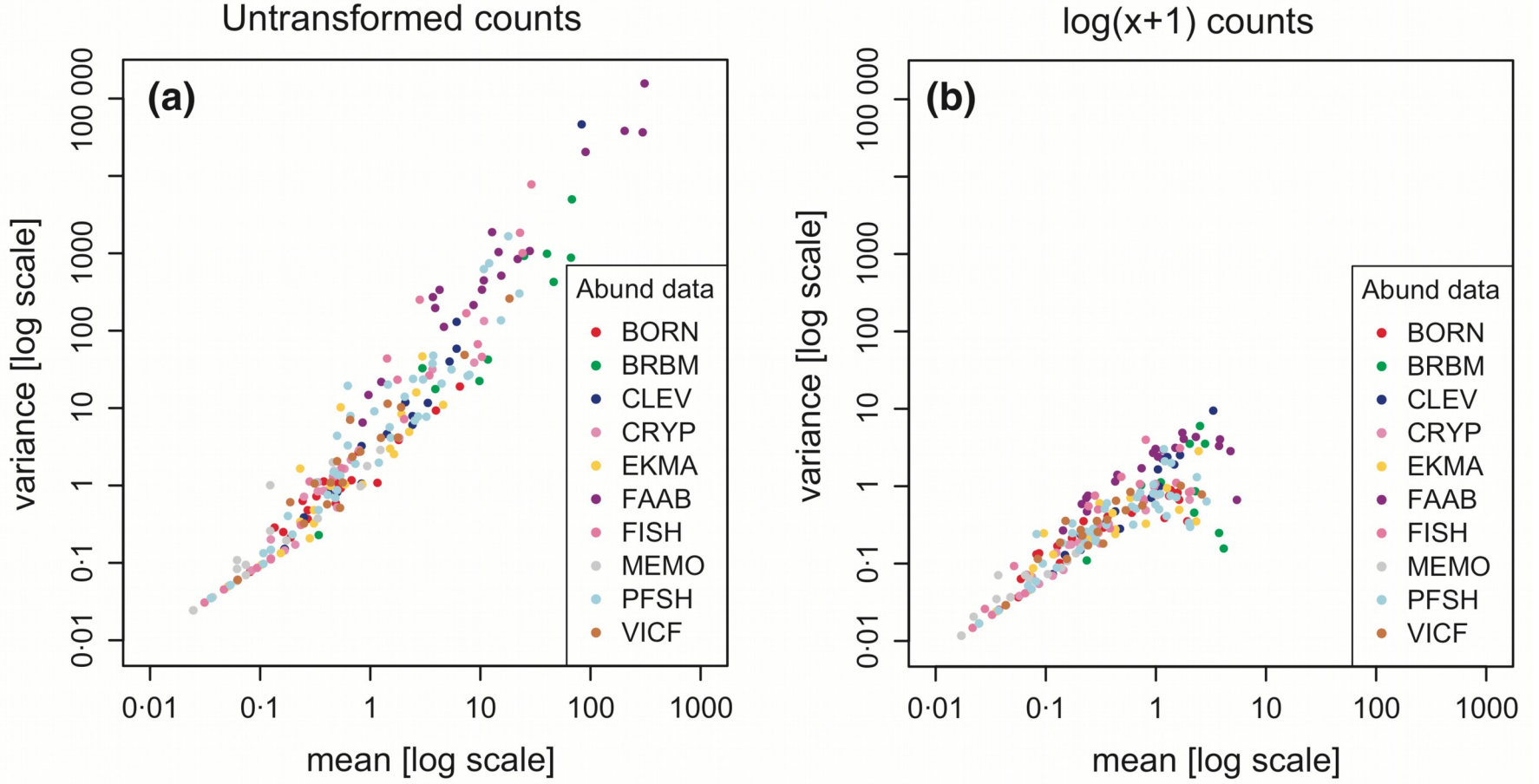
- Dominant approach in last decades
- Computationally very efficient
- Primarily applicable to research goals of explanation, exploration and assessing hypotheses and determining probabilities

Dissimilarity- and algorithm-based ordination



Criticism of dissimilarity- and algorithm-based ordination

- Lack of probabilistic framework (e.g. statistical data model)
- Less interpretable due to lack of parameters at level of observations
- Ignorance of mean-variance relationship



Warton et al. 2011 *Meth. Ecol. Evol.* 3: 89

Model-based ordination

Y = Response (sites x species) transformed or non-transformed

X = Predictors (sites x environmental vars.)

Unconstrained analysis (e.g. GLLVMs, UAO)

Constrained analysis (e.g. multivariate GLMs, CAO)

- Model fitting for each response

$$g(\mu) = \beta_0 + \ldots$$

- Model diagnosis

$$\text{Var}(Y) = \phi\, V(\mu)$$

- Inference based on individual responses

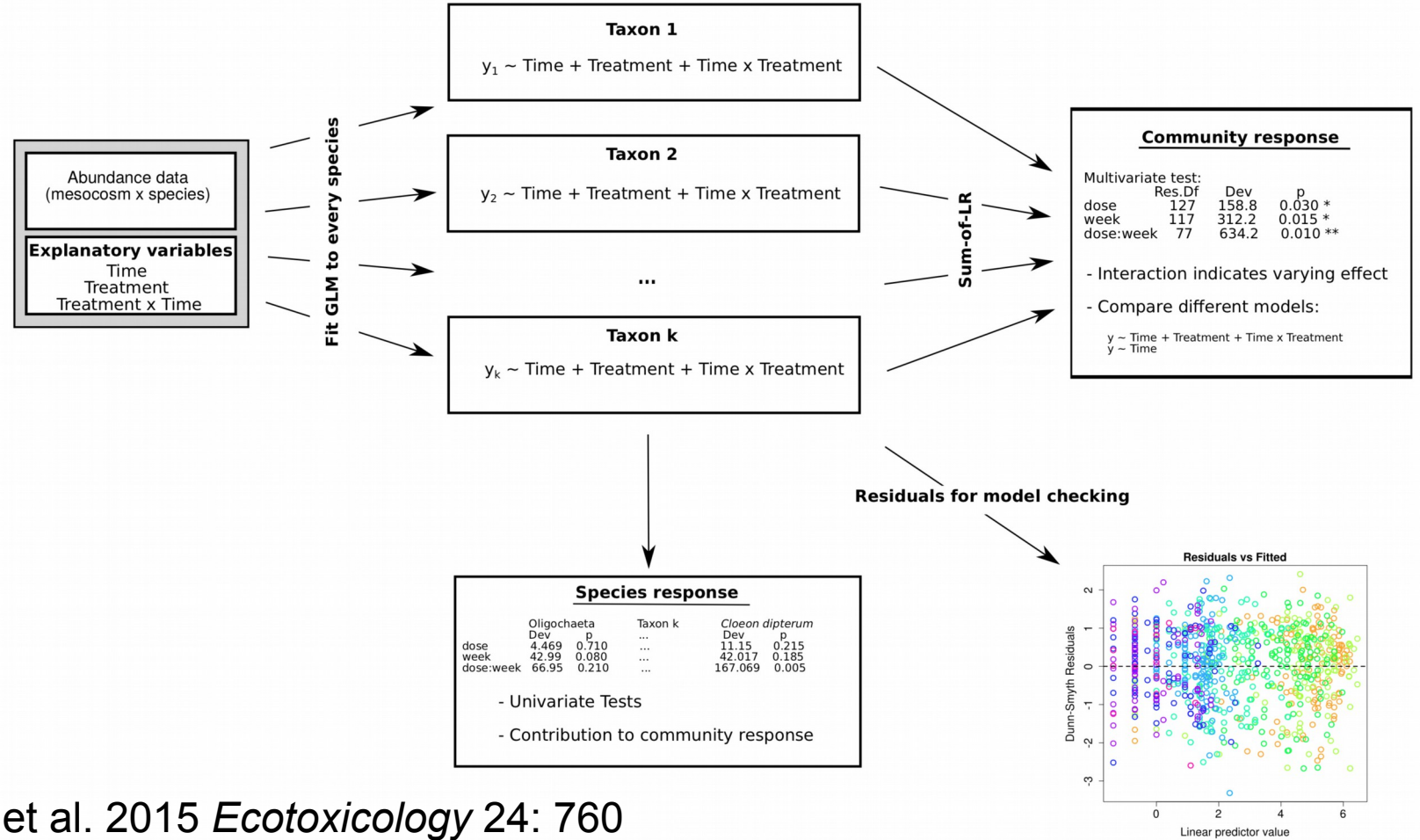
Sum of LR-statistic, $\mathbf{B}_{(-1)} = \mathbf{C}\mathbf{A}^T$

- For several models: Ordination biplot
- Parameter estimates with confidence intervals
- Often likelihood statistics for model

- Emerging approach
- Computationally heavy (not feasible in the last century)
- Applicable to a wide range of research goals
- Explicitly defines signal (systematic model component) and noise (assumed distribution of error)

Multivariate GLMs (*sensu* Warton et al. 2011)

- Fit GLM for each response (note analogy to RDA)
- Diagnose model (e.g. check mean-variance assumption)
- No ordination biplot, mainly for assessing hypotheses, prediction and parameter estimation
- Can fit model analogous to PRC



Szöcs et al. 2015 *Ecotoxicology* 24: 760

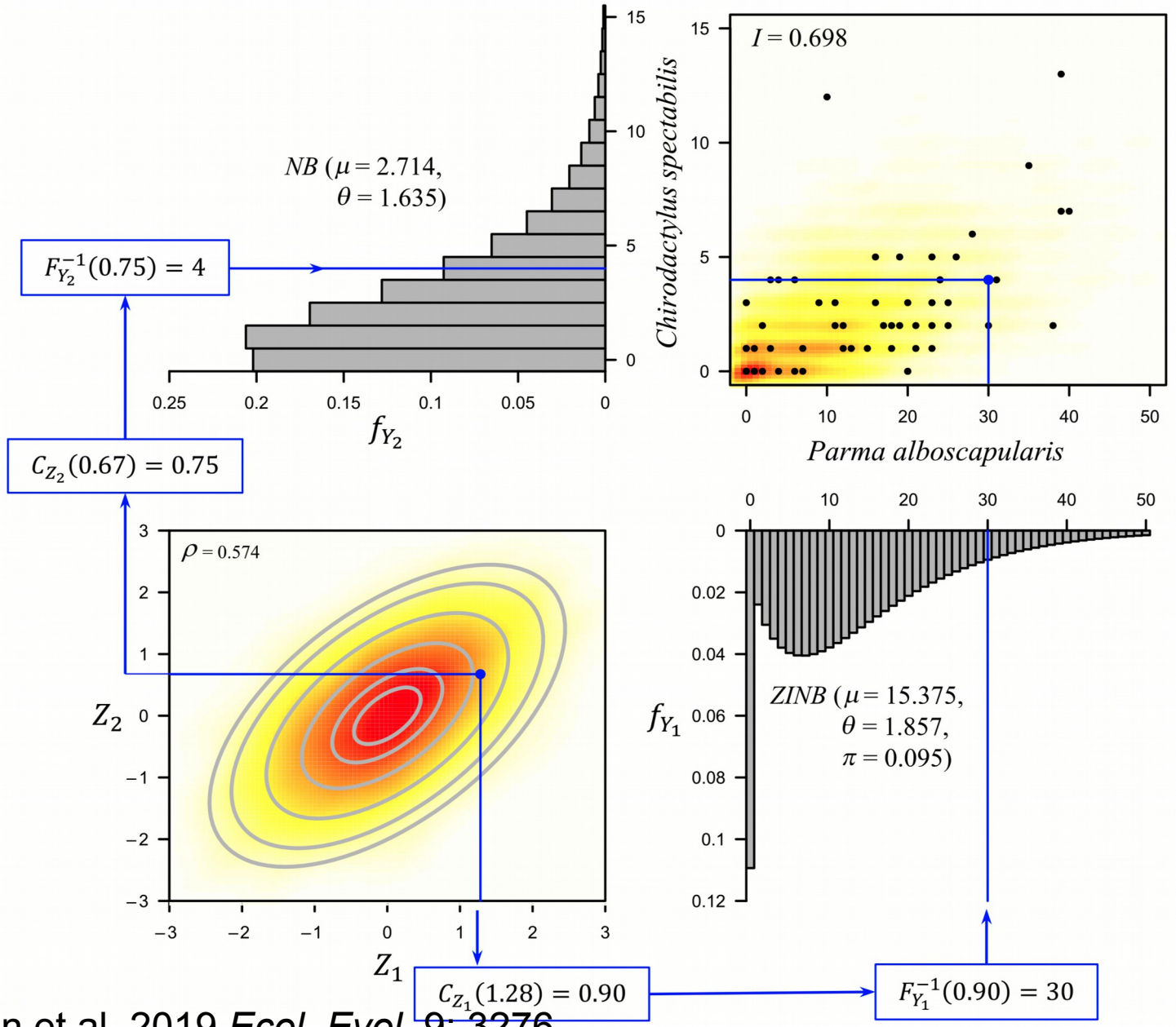
Multivariate GLMs and more complex models

How to account for the correlation between responses (e.g. species) in multivariate GLMs and other models?

- Estimate correlation from data and consider in model → Sample Variance-Covariance matrix $\boldsymbol{S}$ (unless true matrix $\boldsymbol{\Sigma}$ known). Only reliable if few responses or $n >>$ number of responses.
- Assume no correlation (i.e. $\boldsymbol{\Sigma} = \boldsymbol{I}$). In most cases assumption wrong, but reliable p-values can be obtained through design-based inference (based on permutation)
- More complex procedures: Dimension reduction of $\boldsymbol{\Sigma}$ via latent variables, Generalized estimation equations, Copulas.

Copulas for modelling species interactions

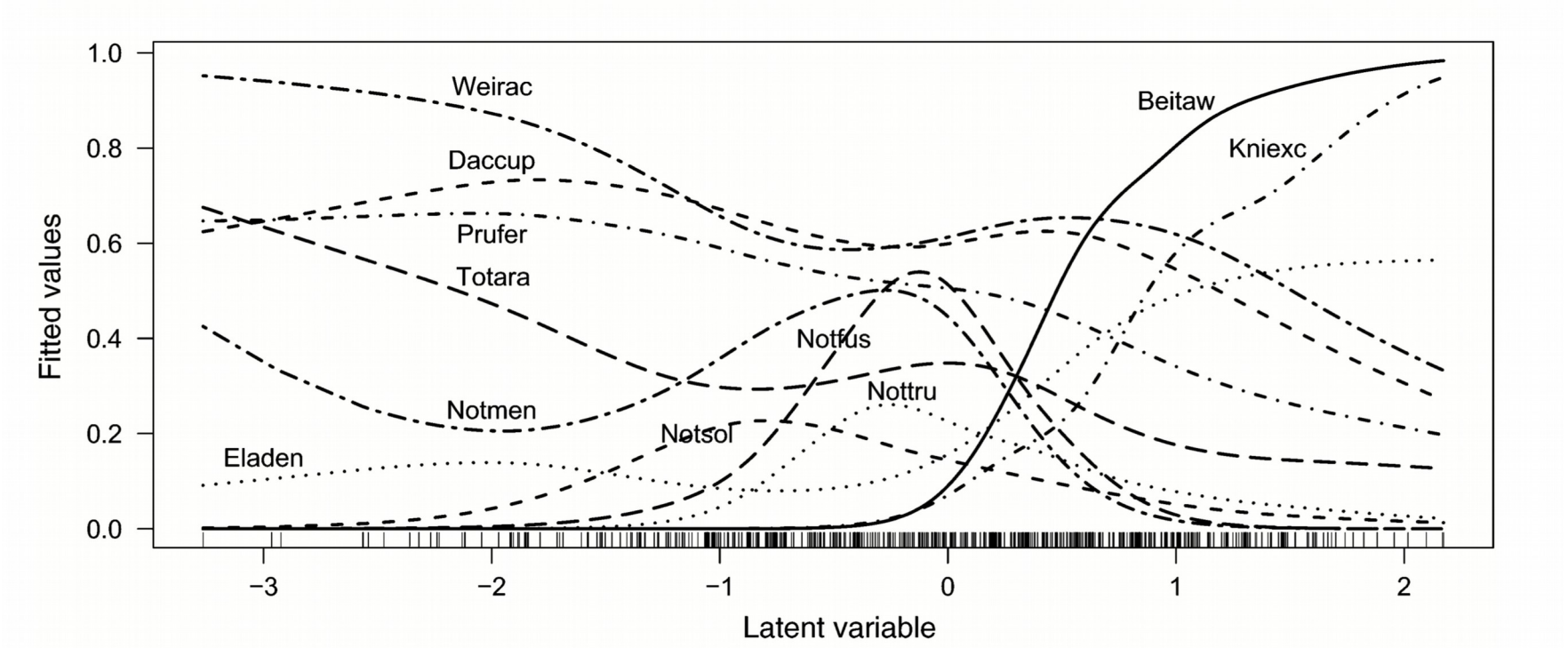
Copula: Function representing joint distribution based on the cumulative distribution functions of individual variables → can couple variables



 Anderson et al. 2019 *Ecol. Evol.* 9: 3276

Constrained Additive ordination

- Data-based rather than model-based ordination: Relies on Generalized Additive Models $g(\mu) = \beta_0 + f(v)$
- Derives response of each species to main environmental gradient from data → no linear or unimodal model assumed, currently restricted to one latent variable
- Computationally very demanding
- Example: Several species deviate from linear and unimodal shape



 Yee 2006 *Ecology* 87: 203

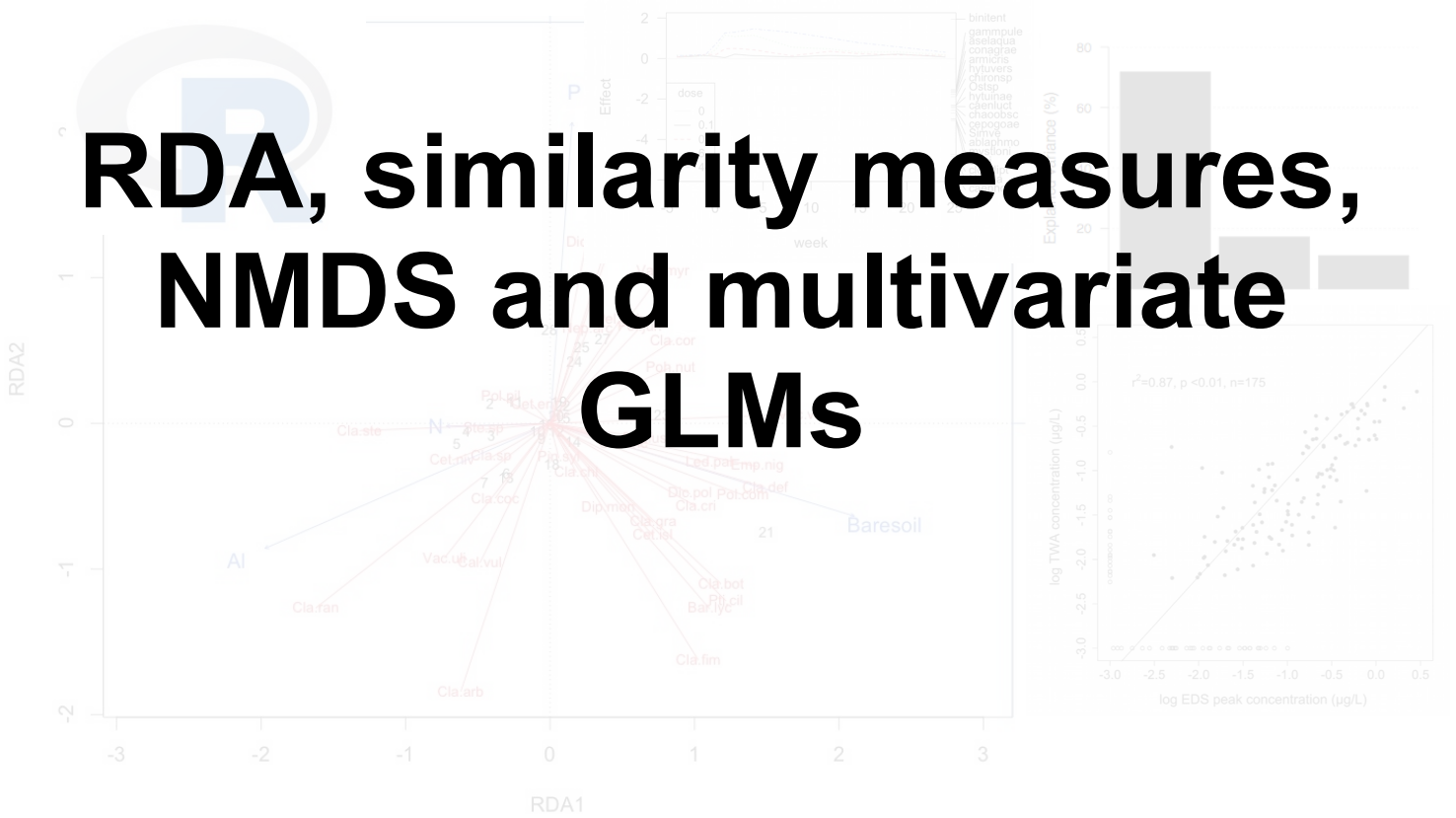
Goodbye, dissimilarity-based methods?

- Comparisons between methods paint a more nuanced picture, model-based approaches not always superior
- Joint approaches (e.g. Anderson et al. 2019)
- Mind your data properties (e.g. mean-variance relationship, response shape) and research goal/question!

https://www.needpix.com/photo/download/935292/puppy-dogs-collie-cute-pet-sweet-free-pictures-free-photos-free-images

Tools for complex data analysis

University of Koblenz-Landau 2020/21



RDA, similarity measures, NMDS and multivariate GLMs

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

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

Learning targets

- Understanding the basics of RDA and extensions.
- Knowledge on the calculation of commonly used association measures.
- Understanding the mathematical background and how to conduct a NMDS.
- Knowledge on model-based ordination and multivariate GLMs

2

Learning targets and study questions

- Understanding the basics of RDA and extensions.
  - How many constrained axes has an RDA and how are they related to the descriptors?
  - What can be interpreted in a RDA triplot?
  - How is PRC related to RDA and for which research contexts is it useful?
- Knowledge on the calculation of commonly used association measures.
  - Which association is measured with similarity measures?
  - For which data are the Bray-Curtis and the Jaccard coefficient suitable?
  - How can similarity measures be visualised?

3

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 types of matrices are computed during NMDS?
  - Outline the main steps when computing a NMDS.
  - Discuss limitations of NMDS.
- Knowledge on model-based ordination and multivariate GLMs
  - Discuss the limitations of dissimilarity-based and algorithm-based methods. How do model-based methods overcome these limitations?
  - Explain the basics of multivariate GLMs.

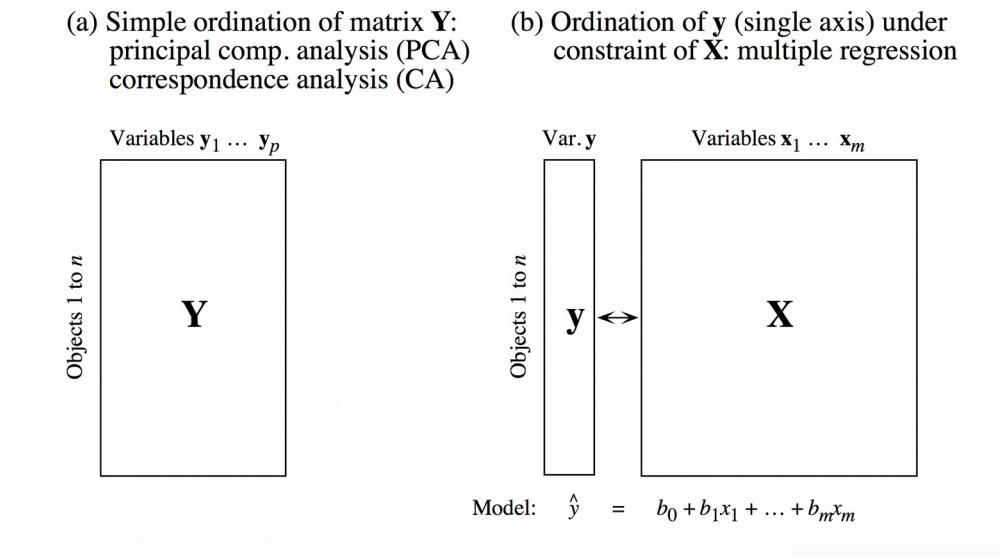
4

RDA, similarity measures, NMDS and multivariate GLMs

Contents

5

Constrained ordination methods

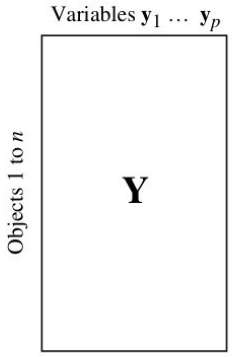


Legendre & Legendre 2012: 627

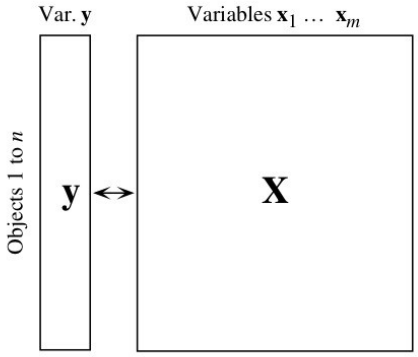
6

Constrained ordination methods

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

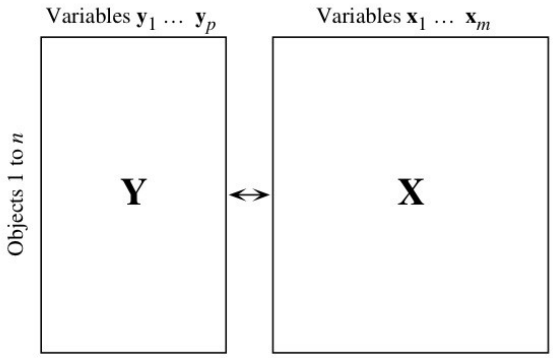


(b) Ordination of **y** (single axis) under
constraint of **X**: multiple regression



Model: $\hat{y} = b_0 + b_1 x_1 + \ldots + b_m x_m$

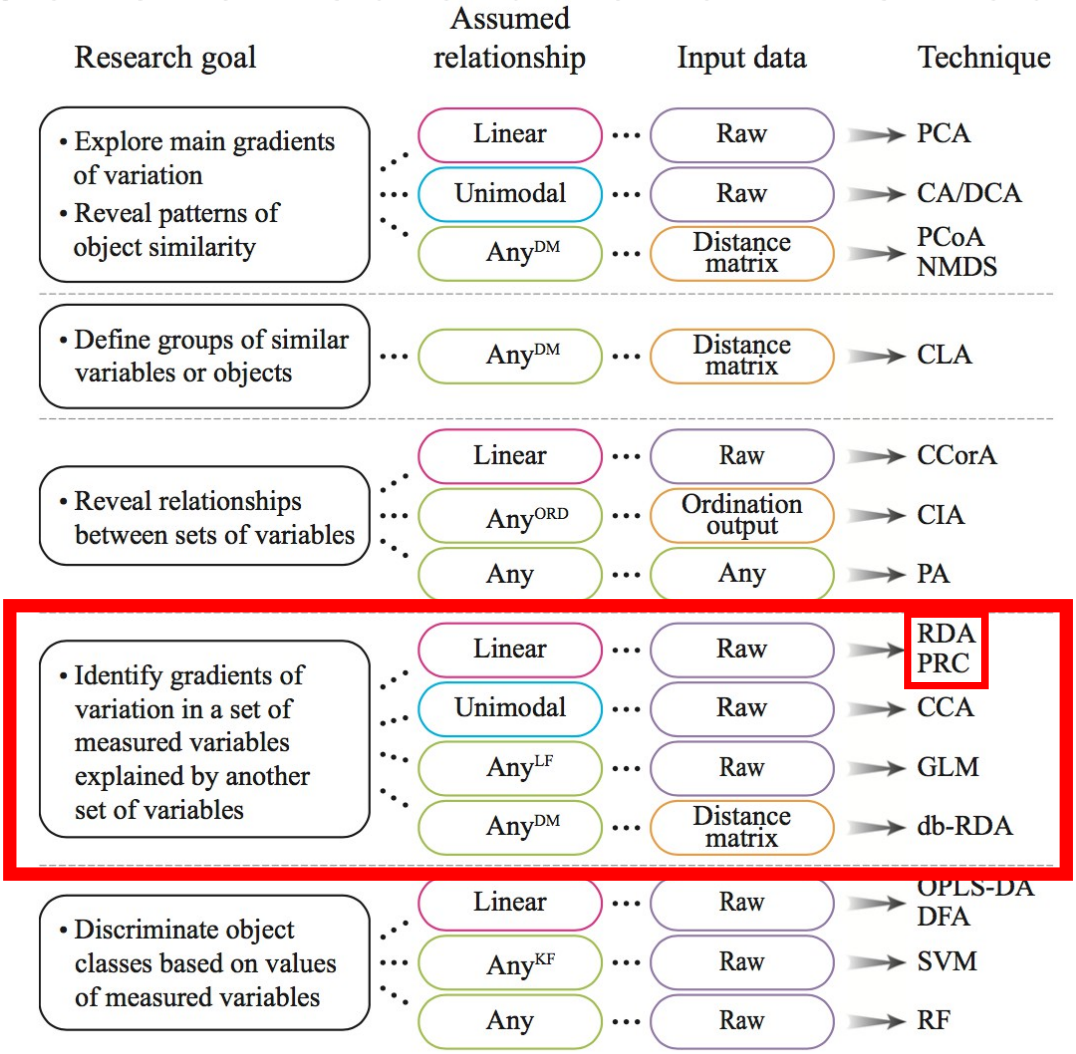
(c) Ordination of **Y** under constraint of **X**:
redundancy analysis (RDA)
canonical correspondence analysis (CCA)



Legendre & Legendre 2012: 627

7

Constrained ordination methods



8

Paliy & Shankar 2016 *Mol Ecol* 25: 1032

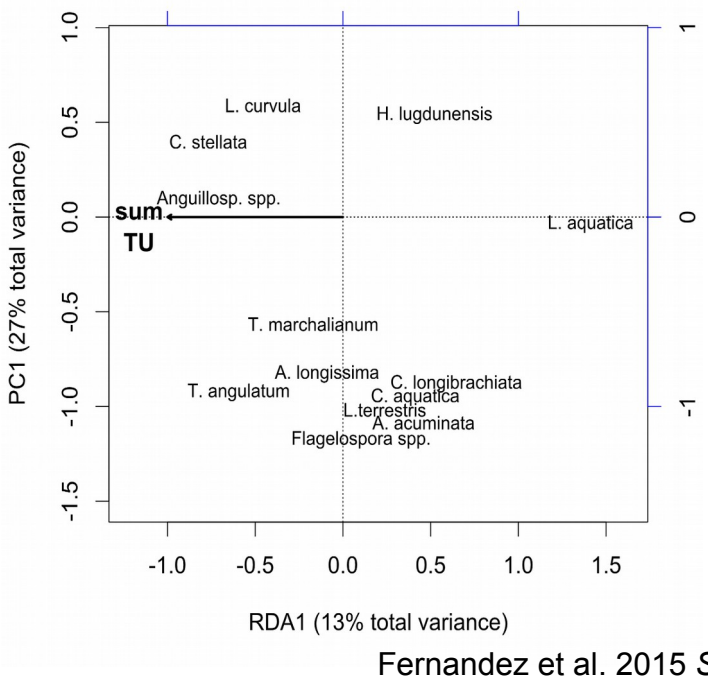
We will discuss RDA and PRC in detail. CCA and db-RDA will be briefly discussed. An alternative approach represent multivariate GLMs, which extend the GLM framework from one response to multiple response variables. The differences between these approaches and the multivariate GLMs will be discussed in the last part of this session.

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

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?



Fernandez et al. 2015 *Sci. Tot. Environ.* 533: 40

Redundancy = explained variance

9

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

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

$$\hat{y} = \boldsymbol{X} b$$

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

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

Reformulation for multivariate multiple regression with several y:

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

10

Note that in previous lectures we used the matrix notation for y in the equations, i.e. $\boldsymbol{Y}$ where $\boldsymbol{Y}$ 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 $n \times m$ response matrix $\boldsymbol{Y}$ for the m response variables y.

Mathematical background of RDA

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

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

Usually, this is not known and the sample variance-covariance matrix $\boldsymbol{S}$ (also called Dispersion matrix) is estimated from the observations:

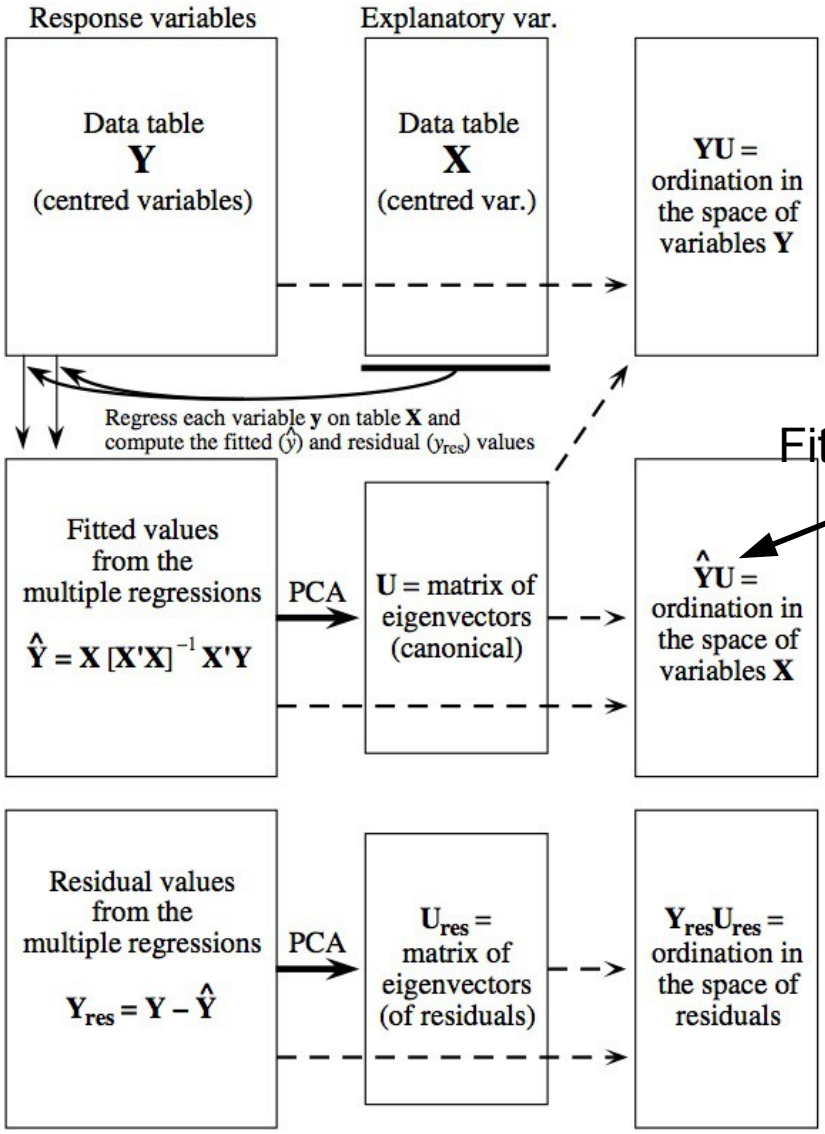
$$\boldsymbol{S}_{\hat{\boldsymbol{Y}}^T\hat{\boldsymbol{Y}}} = \frac{1}{n-1}\hat{\boldsymbol{Y}}^T\hat{\boldsymbol{Y}}$$

and used in a PCA: $\boldsymbol{S}_{\hat{\boldsymbol{Y}}^T\hat{\boldsymbol{Y}}}\,a = \lambda\,a$ — Eigenvector

Eigenvalue problem

➡ Eigenvectors linear combinations of predictors

11

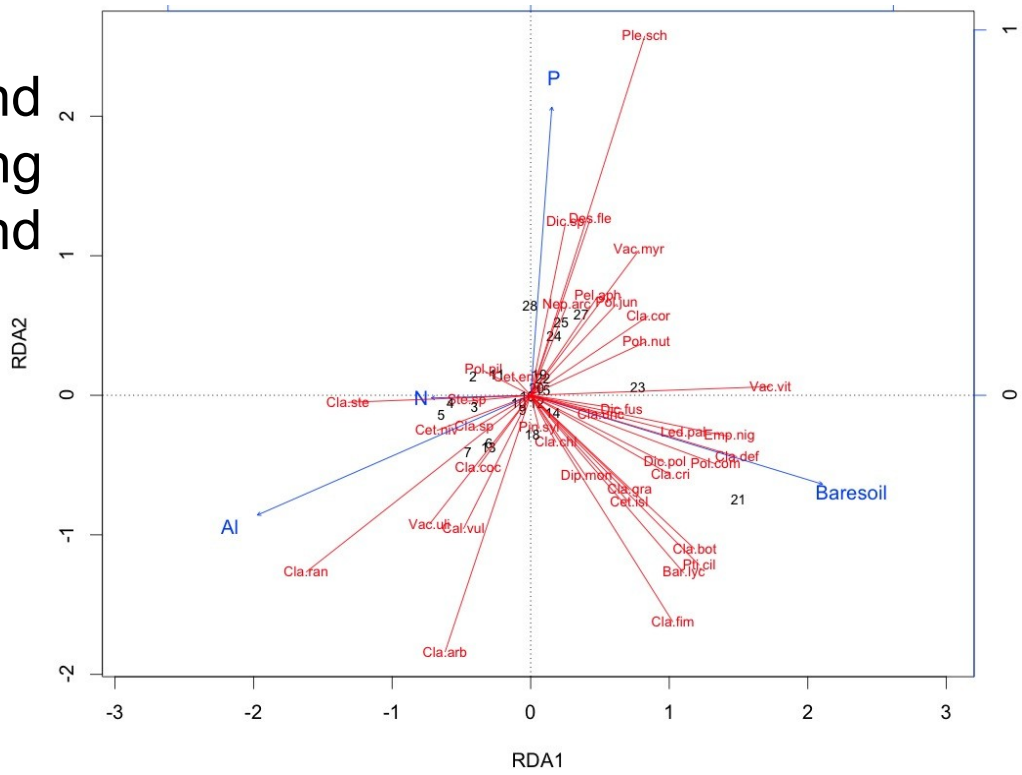


12 Legendre & Legendre 2012:631

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



We will discuss the RDA results in detail in the R tutorial. The issue of scaling has been extensively discussed for PCA.

13

RDA, similarity measures, NMDS and multivariate GLMs

Contents

1. Constrained ordination and RDA
2. **RDA assumptions and PRC**
3. Similarity measures
4. Nonmetric multidimensional scaling (NMDS)
5. Model-based ordination, Multivariate GLMs

14

RDA model and variable importance

How many RDA axes are required?

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

Which explanatory variables should be included in the best-fit RDA if the research goal is explanation and how important are they?

- Manual and automatic model-building with adjusted R^2 as goodness of fit criterion (as for multiple linear regression) and permutation-based p-values
- Variance partitioning between different models to determine explained variance of individual explanatory variables

Remember our discussion of the modelling strategies in the context of the linear model, where we emphasised the role of the research goal. In the case of testing scientific hypotheses or parameter estimation, you would typically have pre-specified (RDA) models. Similarly, for prediction you would typically include all scientifically relevant variables and use the full (RDA) model (see Session 5 for further details). Model-building is mainly relevant if the research goal is explanation. For details on model building for RDA as well as partitioning of variance see Borcard et al. 2018: 221-238.

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 (objects) >> p (predictors/explanatory variables) to reliably infer p importance
- RDA can be employed for multivariate ANOVA (see Borcard et al. 2018: 238 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)

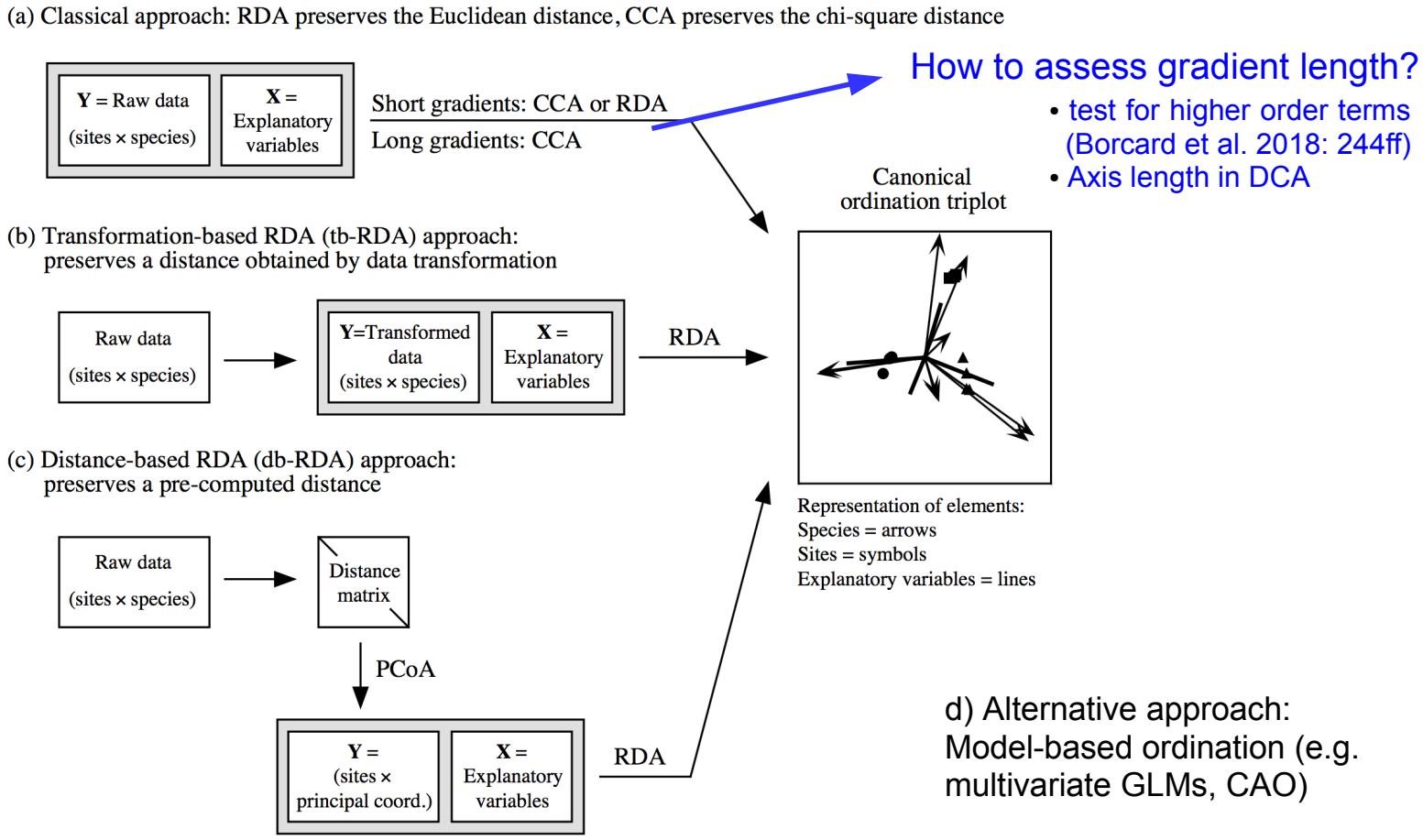
16

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 the 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.

A multivariate ANOVA (also MANOVA) is an analysis of variance, where the response is a quantitative matrix $\boldsymbol{Y}$ (e.g. concentrations of several minerals in soil), constituted by multiple response variables y (cf. discussion of Mathematical background of RDA).

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 extensions



17 Legendre & Legendre 2012: 648

DCA: Detrended Correspondence Analysis
CCA: Canonical Correspondence Analysis

Transformation-based RDA is discussed in Legendre and Gallagher (2001). Information on db-RDA can be found in Legendre & Anderson (1999) and McArdle & Anderson (2002). As an example for a study where db-RDA is applied see Szöcs, Kefford and Schäfer (2012).

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

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

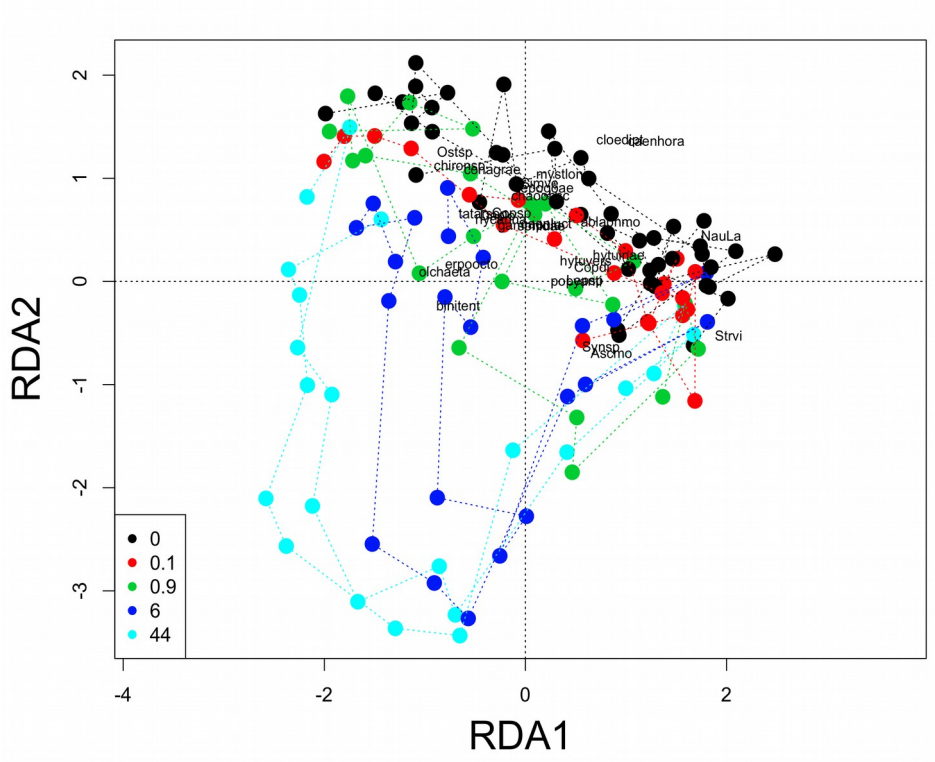
McArdle, B. H. & Anderson, M. J. (2002) Fitting multivariate models to community data: A comment on distance-based redundancy analysis. *Ecology* 82 (1), 290-297.

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 macroinvertebrates? *Sci. Total Environ.* 437 (1), 121-126.

RDA extension: Principal Response Curve (PRC)

Example: Before-After-Control-Impact (BACI) study with communities. Treatment of aquatic mesocosms containing invertebrates with insecticide chlorpyrifos.

- RDA model for time, treatment and their interaction
- Clear time and treatment effect but figure cluttered
- Pure time effect often not relevant → Remove from model



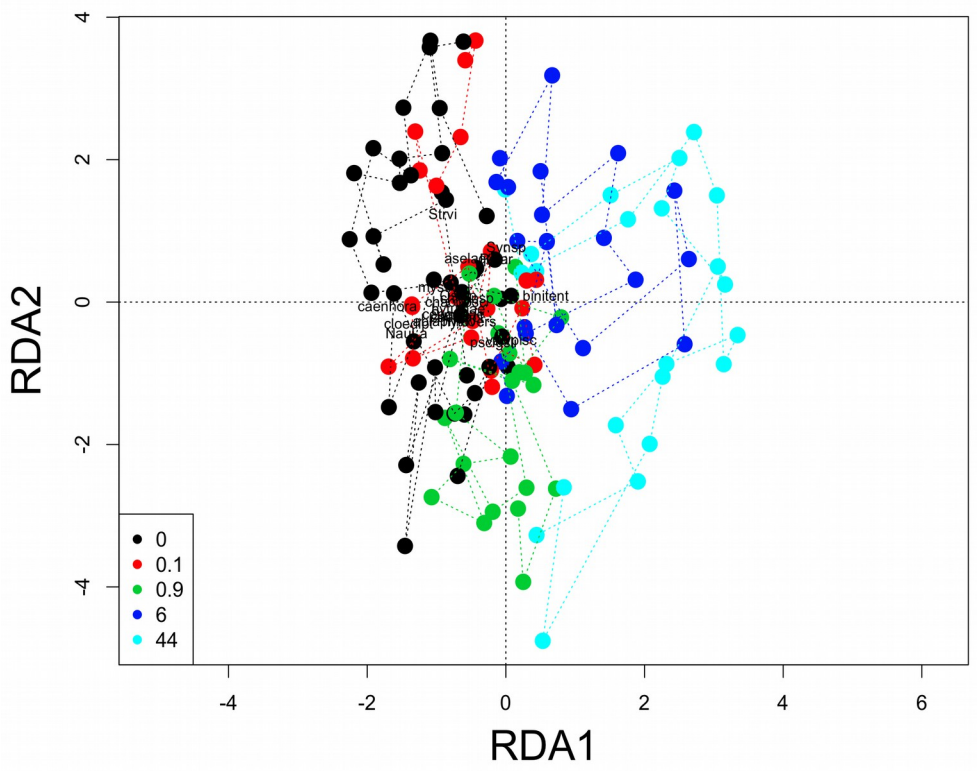
In previous versions of this course, the PRC has been featured more extensively. It is a method that is specifically relevant for ecotoxicological data analyses, for example of community data from mesocosm experiments. The material from the previous version of the lecture can be accessed here: https://github.com/EDiLD/permanova_lecture/tree/master/prc

18

RDA extension: Principal Response Curve (PRC)

Partial RDA: Explanatory variables held constant to remove their effect (Borcard et al. 2018: 221-225)

- Time effect removed ("partialled out") with partial RDA
- First axis: Treatment and interaction effect
- Better separation between treatments, still cluttered
  → Focus on 1st axis

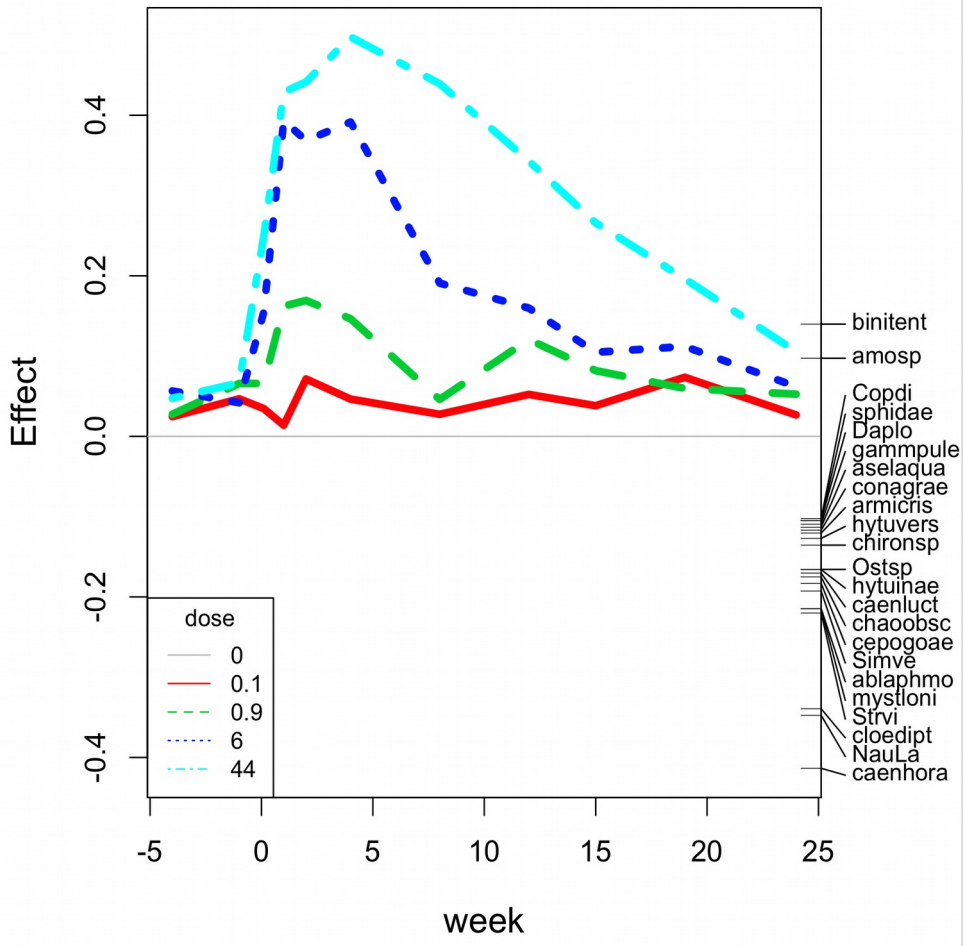


We have discussed Added-variable plots that are also called partial regression plots. These are described in detail in the R tutorial http://139.14.20.252:3838/session/5/#section-model-visualisation. Partial RDA follows a similar approach as partial regression, where first the effect of the other variables is removed by fitting a model and then proceed by using the residuals from this model (which represents the variance not explained by the conditioning variables – this is called "partialling out their effect").

19

Principal Response Curve (PRC)

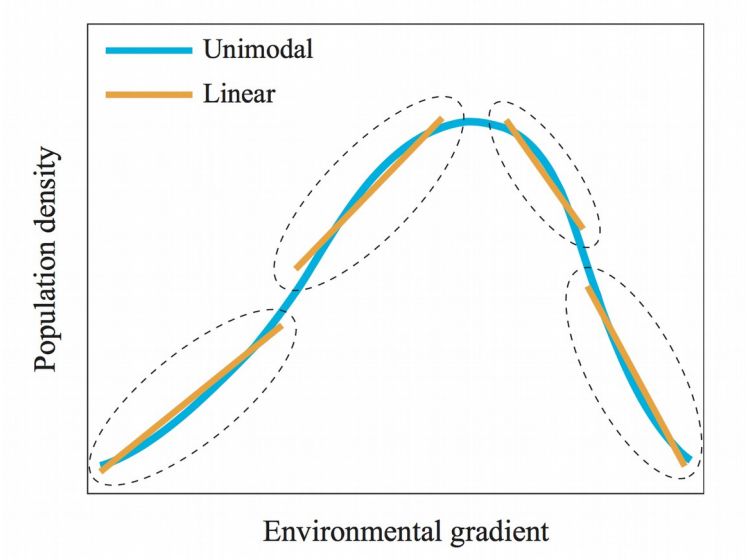
- 1st axis of partial RDA
- Treatments plotted relative to control by subtracting site scores
- Y axis: Treatment effect (difference in composition)
- X axis: Time
- Species scores: Species responsible for pattern
- Recovery if treatments approach control



The PRC displays the 1st axis of the partial RDA where time is "partialled out". Thus, it represents the effect of treatment and of the interaction of treatment with time.

Low species scores do not mean that the taxa was non-responsive to the treatment. If the response pattern of a taxon differs from the general pattern of the PRC (in the case study, for example, a species that does not recover would differ from the general pattern), the species score would typically be low, although it may have responded to the treatment.

Alternative to RDA for unimodal responses



Canonical Correspondence Analysis (CCA)

- Similar to RDA, but assumes unimodal distribution (χ^2-distance) of species along environmental gradient → widely used for ecological data
- Extension of (unconstrained) correspondence analysis
- Similar modelling framework as for RDA

21

Paliy & Shankar 2016 *Mol Ecol*:1032

We touch only briefly on CCA. We discussed before how to diagnose gradient length. 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)).

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

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

RDA, similarity measures, NMDS and multivariate GLMs

Contents

1. Constrained ordination and RDA
2. RDA assumptions and PRC
3. **Similarity measures**
4. Nonmetric multidimensional scaling (NMDS)
5. Model-based ordination, Multivariate GLMs

22

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

23

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 for occurrence data

Simple matching coefficient

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

$$S_{\text{Match}} = \frac{a+d}{a+b+c+d}$$

Exercise: Calculate S_{Match} for the data below with and without the 1. and 4. species. How do these species influence S_{Match}?

Site				
1	0	400	0	0
2	0	0	10	0

Occurrence data is also termed presence-absence data and represents a case of binomial data.

Similarity measures for occurrence data

Site				
1	0	400	0	0
2	0	0	10	0

$$S_{\text{Match}} = \frac{a+d}{a+b+c+d}$$

Calculation with all species:

a = 0, b = 1, c = 1, d = 2 → S_{Match} = 2/4 = 0.5

Calculation without species 1 and 4:

a = 0, b = 1, c = 1, d = 0 → S_{Match} = 0/2 = 0

Joint absence of species influences similarity between sites

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

→ **Double-zero problem**

25

For details on the double-zero problem see Legendre & Legendre (2012: 271-272).

Widely used similarity measures

Jaccard coefficient S_{Jacc} (=Jaccard similarity index)

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

$$S_{\text{Jacc}} = \frac{a}{a+b+c}$$

- occurrence data
- ignores joint absences (d)
- Range: 0 (no similarity) to 1 (identity)

Bray-Curtis coefficient S_{BC}

- abundance data
- Range: 0 to 1
- Often prior data transformation to reduce weight of dominant taxa

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

$x_{i,k}$ and $x_{j,k}$ is the abundance of taxon k in site i and j.

26

For non-negative input data, which is obviously the case for species abundance or occurrence data, the coefficients range from 0 to 1.

Similarity measures 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).

Dissimilarity measures represent the counter part to similarity measures. A similarity measure S can be converted to the corresponding dissimilarity measure D via: $D = 1 - S$. The Jaccard dissimilarity D_{Jacc} is given as: $D_{\text{Jacc}} = (b+c)/(a+b+c)$ or 1 - S_{Jacc}. S_{Jacc} (and D_{Jacc}) 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: Transformation and S_{BC}

Site				
i = 1	0	400	5	0
j = 2	0	0	10	0
Min	0	0	5	0
Sum	0	400	15	0

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

Calculation:

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

Calculation for square-root transformed data:

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

Calculation for double square-root transformed data:

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

The dissimilarity D_{BC} for Bray-Curtis is 1-S_{BC} and can be calculated for two sites i and j 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, for example due to a chemical spill, a dummy species can be added to the data to enable calculation of S_{BC} (Clarke et al 2006).

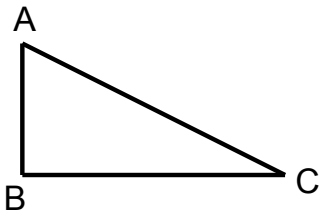
Note that the log (x+1) transformation increases the weight of rare taxa, whereas square-root or double-square root transformation does not. For a discussion on transformations for ecological data see Legendre & Gallagher (2001).

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.

Distance measure

Association (e.g. dissimilarity) measure meeting triangle inequality criterion:

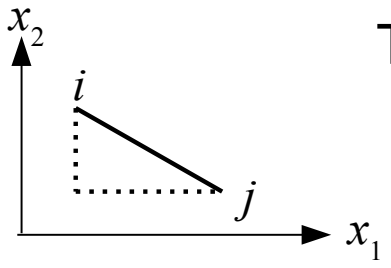


d(A,B) + d(B,C) ≥ d(A,C), where d is distance function
Sum of any two sides of triangle always ≥ third side

Important for geometrical representation (e.g. ordination)

<u>Euclidean distance:</u> frequently used distance measure (e.g. PCA, RDA), not suitable for ecological data (→ species abundance paradox)

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



Two dimensions:

$$D_{\mathrm{Eucl}}(i,j) = \sqrt{(x_{i,1}-x_{j,1})^2+(x_{i,2}-x_{j,2})^2}$$

28

The definition of the distance measure follows that of Everitt et al. (2011: 49). By contrast, Legendre & Legendre (2012) do not reserve the term distance measure for dissimilarity measures that meet the triangle inequality criterion. Several dissimilarity measures can be square-root transformed to meet the triangle inequality criterion, see Legendre & Legendre 2012: 295-297 for details. Moreover, several distance measures have no corresponding similarity measure (e.g. Euclidean distance, Hellinger distance).

The species abundance paradox has been discussed in the session on PCA. We also discussed the Mahalanobis distance that can be used to evaluate the distance between two multivariate vectors (objects in the terminology used in the context of association measures) or between individual multivariate vectors (objects) and the overall mean vector, taking the covariance among descriptors into account, which renders the Mahalanobis distance independent from different scales in the data set. If the covariance matrix for two multivariate vectors is the identity matrix, then the Mahalanobis distance is equivalent to the Euclidean distance D_{Eucl}. Conversely, if the scales of the descriptors vary (covariance matrix differs from identity matrix), D_{Eucl} is biased towards the descriptors with larger scales. For example, in the case of species data, D_{Eucl} is typically dominated by the species with the largest absolute difference in abundance, unless the magnitude of absolute differences are similar between all species.

Everitt et al. 2011 *Cluster Analysis*. John Wiley & Sons.

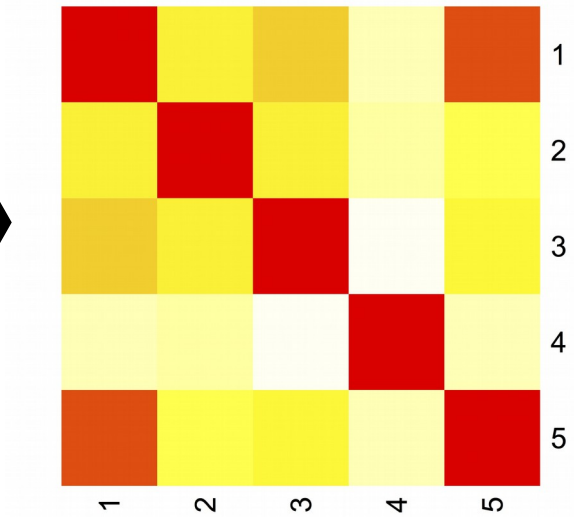
Visualization of association measures

Heatmap

- Associations converted to colours
- Relationship easier to grasp

Matrix of distances between sites

	1	2	3	4	5
1	0	0.69	0.6	0.92	0.22
2	0.69	0	0.7	0.89	0.8
3	0.6	0.7	0	0.98	0.72
4	0.92	0.89	0.98	0	0.92
5	0.22	0.8	0.72	0.92	0



Diagonal entries: Distance of site to itself (= 0)

29

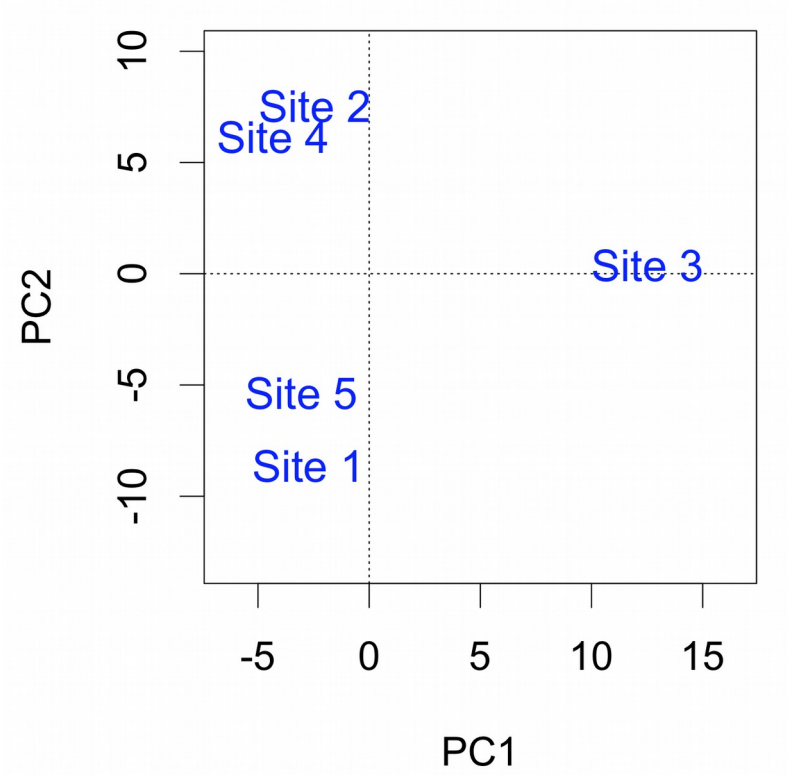
For an overview of techniques to visualize multivariate ecological data see:

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

Visualization of association measures

Ordination

Measures that meet triangle inequality criterion allow for clear geometrical interpretation of ordination



30

How to select an association 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: *simple matching* (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: *Euclidean distance* (D_1) and *average distance* (D_2)

31

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

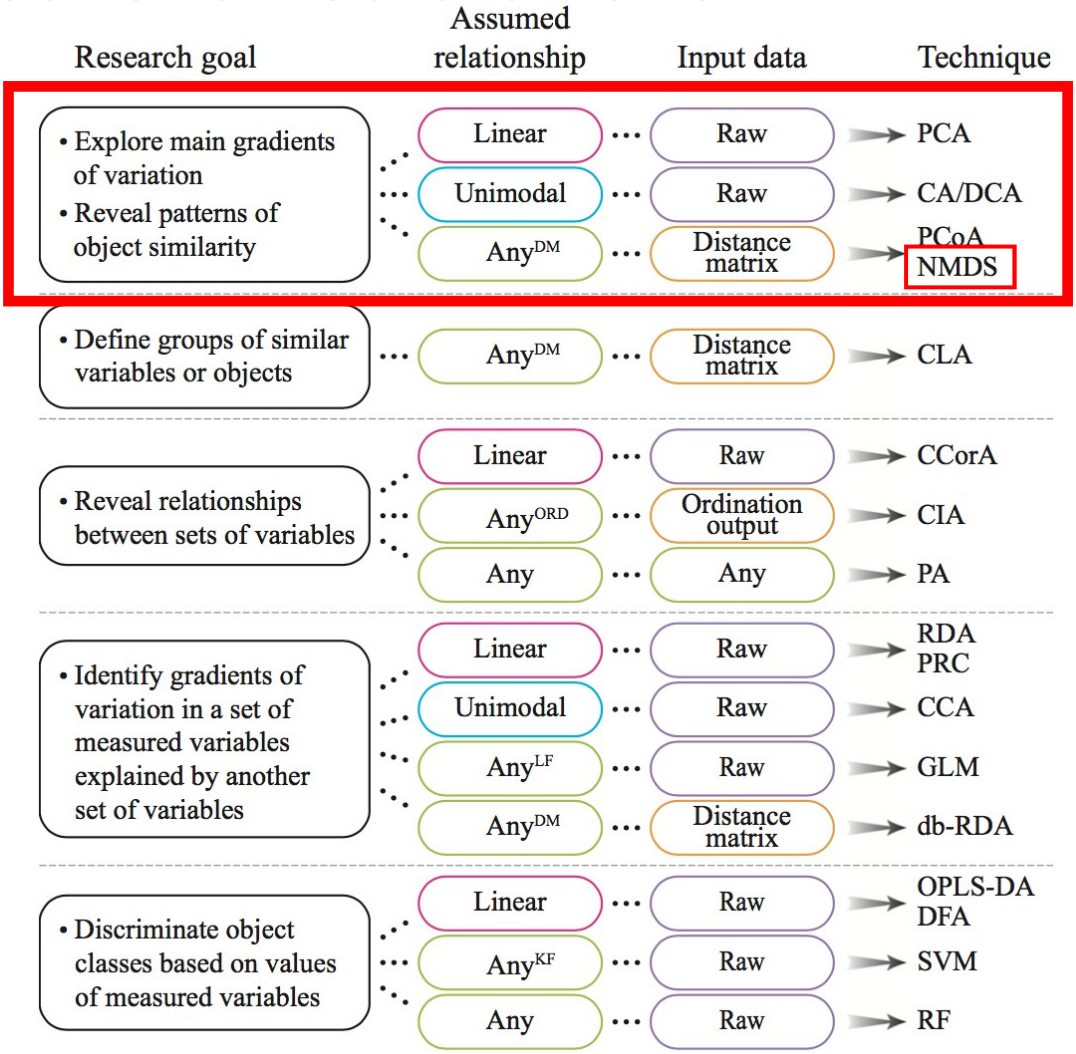
RDA, similarity measures, NMDS and multivariate GLMs

Contents

1. Constrained ordination and RDA
2. RDA assumptions and PRC
3. Similarity measures
4. **Nonmetric multidimensional scaling (NMDS)**
5. Model-based ordination, Multivariate GLMs

32

Unconstrained ordination with NMDS



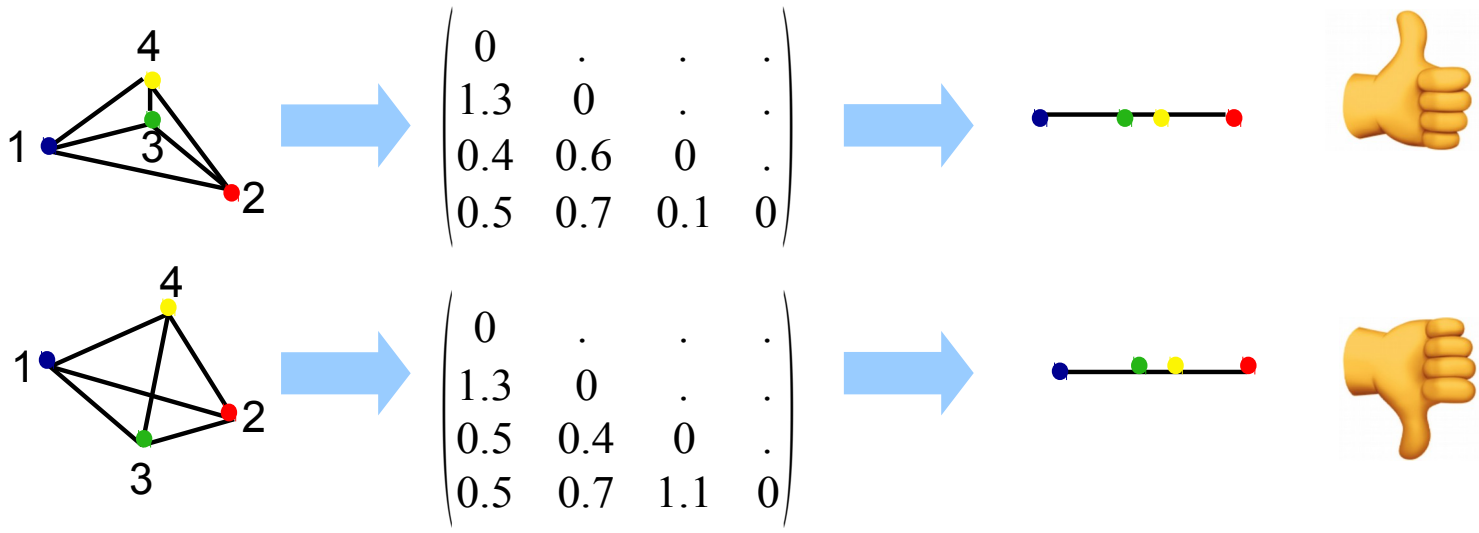
33

Paliy & Shankar 2016 *Mol Ecol*:1032

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

Challenge of lower-dimensional preservation of metric distances

Example for one-dimensional representation of two dimensions:

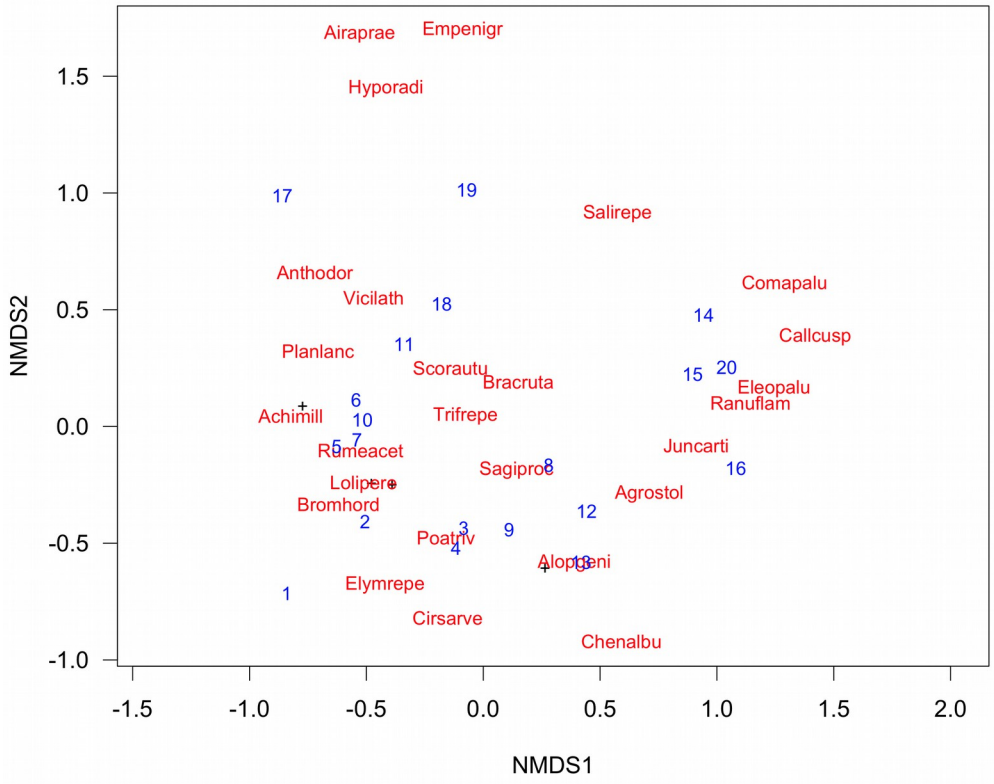


Idea: Relax condition for lower dimensional representation by preserving rank instead of absolute distances

It is easier to preserve ranked instead of absolute distances. In the example, rank-based methods would focus on preserving ranked distances: Highest distance between 1 and 2, second highest between 3 and 4 and so on.

Nonmetric multidimensional scaling

- Unconstrained ordination based on ordered ranks of pairwise distances or dissimilarities (→ nonmetric)
- Can be used with many distance/ dissimilarity measures → Suitable for ecological data
- Not based on eigenvalues, no partitioning of variance
- Very robust and flexible



NMDS is the nonmetric version of multidimensional scaling (MDS). Different algorithms have been developed for MDS, hence the term does rather refer to a group of methods. Generally, MDS aims to achieve a low-dimensional representation of data points such that distances between them approximate the original distances. Metric MDS aims at preserving the original distances between objects, whereas NMDS intentionally relaxes this condition and only aims at preserving the order of the ranked distances. NMDS is therefore more robust to outliers and more easily achieves an appropriate two-dimensional representation of distances, but this comes at the cost of the potential distortion of original distances.

NMDS is regarded as one of the most robust unconstrained ordination methods (Minchin 1987). It is very flexible as it can be used with a variety of distance measures. A recent study showed that NMDS performed reasonably well compared to more advanced methods, unless the data sets were more complex (Roberts 2020).

Minchin, P.R. 1987 An Evaluation of the Relative Robustness of Techniques for Ecological Ordination. *Vegetatio,* **69,** 89-107

Roberts D.W. (2020). Comparison of distance-based and model-based ordinations. *Ecology*. In press. https://doi.org/10.1002/ecy.2908

Steps of NMDS algorithm

1. Determine distance matrix Δ for raw data
2. Set number of dimensions k
3. Set initial configuration
4. Determine distance matrix D for configuration
5. Monotone regression and Pool Adjacent Violators (PAV) algorithm → Disparity matrix $\hat{\mathrm{D}}$ and Goodness of fit measure $\mathrm{STRESS1}$
6. Start with new random configuration and go to 4. (if fit does not improve on many iterations → 7.)
7. Final configuration

36

We will discuss the choice of the number of dimensions k later.

In the case of the R function that we will use, the initial configuration derives from the results of Principal Coordinates Analysis (PCoA). PCoA is sometimes incorrectly described as equivalent to MDS. Although the methods are similar, they minimise different goodness of fit measures. For details on PCoA see Legendre & Legendre (2012: 492-512).

From distance matrix for raw data...

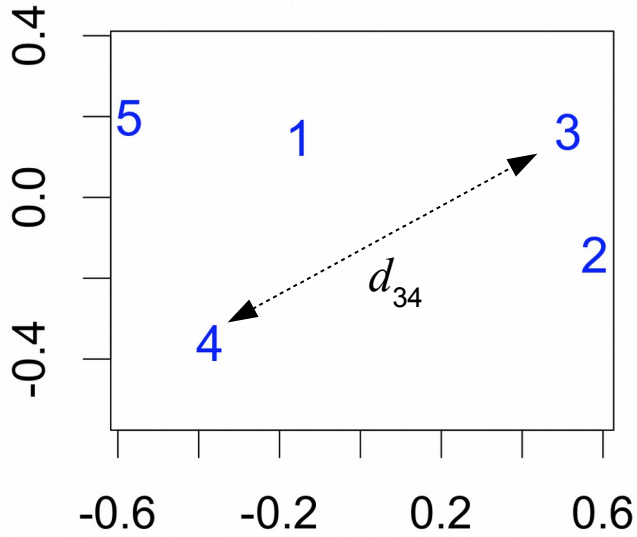
1. Determine distance matrix Δ for raw data

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

Ordered distances of distance matrix

$$\delta_{24} < \delta_{25} < \delta_{23} < \delta_{13} < \delta_{45} < \delta_{35} < \dots$$

2. Set number of dimensions: $k = 2$
3. Set initial configuration



37

Example adopted from Handl (2010).

The number of dimensions needs to be specified before running an NMDS. In the given example we selected two dimensions. Consequently, the initial configuration from PCoA is also for two dimensions.

… over distance matrix for configuration...

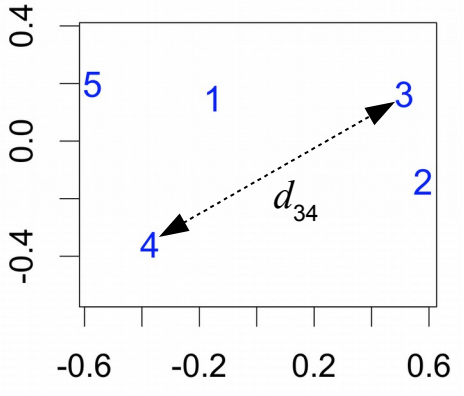
1. Determine distance matrix Δ for raw data

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

Ordered distances of distance matrix

$$\delta_{24} < \delta_{25} < \delta_{23} < \delta_{13} < \delta_{45} < \delta_{35} < \dots$$

4. Determine distance matrix D for configuration



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

$$d_{24} < d_{45} < d_{25} < d_{13} < d_{23} < d_{35} < \dots$$

Order of distances of D not matching with Δ

38

Note that the the relationships between objects in the left figure do not match those of D and the related order of distances, the figure only serves the purpose of illustration.

… to disparity matrix

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

$$d_{24} < d_{45} < d_{25} < d_{13} < d_{23} < d_{35} < \dots$$

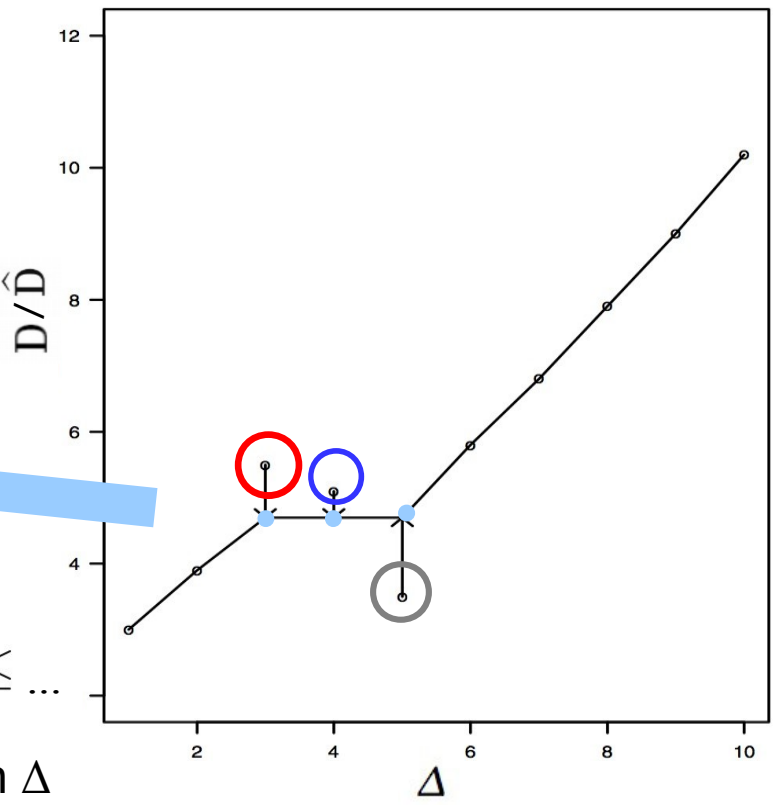
$$\hat{\mathbf{D}} = \begin{pmatrix} 0 & 9.0 & 4.7 & 10.2 & 6.8 \\ 9.0 & 0 & 4.7 & 3.0 & 3.9 \\ 4.7 & 4.7 & 0 & 7.9 & 5.8 \\ 10.2 & 3.0 & 7.9 & 0 & 4.7 \\ 6.8 & 3.9 & 5.8 & 4.7 & 0 \end{pmatrix}$$

$$\hat{d}_{24} \leq \hat{d}_{25} \leq \hat{d}_{23} \leq \hat{d}_{13} \leq \hat{d}_{45} \leq \hat{d}_{35} \leq \dots$$

Order of distances of $\hat{\mathbf{D}}$ matching with Δ

$$\delta_{24} < \delta_{25} < \delta_{23} < \delta_{13} < \delta_{45} < \delta_{35} < \dots$$

5. Monotone regression and PAV algorithm



Handl 2010: 174

The matrix $\hat{\mathbf{D}}$ is called disparity matrix. The disparity matrix is obtained by conducting a monotone regression analysis on the configuration distance matrix. In the case of NMDS, the distances of the configuration should not decrease along the X axis that represents the original distances. The PAV algorithm searches for objects that violate this monotonicity assumption and averages the distances of these objects (for details see Härdle & Simar (2015: 466-471)). The disparity matrix stores the results of the monotone regression analysis with PAV algorithm. Consequently, the distances in the disparity matrix meet the monotonicity assumption.

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

Goodness of fit and number of dimensions

- Goodness of fit metric STRESS1: Difference between original distance and distance of final configuration (in $\hat{D}$)

$$\text{STRESS1} = \sqrt{\frac{\sum_{i<j} (d_{i,j} - \widehat{d_{i,j}})^2}{\sum_{i<j} d_{i,j}^2}}$$

- Rules of thumb for interpretation (Clarke 1993)

Value of STRESS1	Goodness of configuration
< 0.05	excellent
< 0.10	good
< 0.2	medium
> 0.2	bad

- Problem: STRESS1 dependent on factors such as data type, sample size, dimensionality.
  → Alternative: Permutation-based assessing of hypotheses (Dexter et al. 2018)

Which number of dimensions to set?

- Main purpose of NMDS is visualisation: 2-3 dimensions
- Use thresholds related to STRESS1 (but see Dexter et al. 2018)

40

Dexter et al. (2018) provide R code to run a permutation-based hypothesis test for a specific data set.

Typically, 2-3 dimensions are used for visualisations because interpretation becomes cumbersome with more dimensions.

Regarding STRESS1 the threshold 0.2 is often regarded as the maximum value that is acceptable for an NMDS and the number of dimensions is increased if the STRESS1 value exceeds 0.2. Similarly, some authors have selected the number of dimensions that yields to a STRESS1 value below 0.05. However, as discussed for assessing hypotheses based on *p*-values, this dichotomisation of a scale is rather arbitrary, in particular given the dependence of the STRESS1 value on factors such as sample size and data type.

Clarke K.R. (1993). Non-parametric multivariate analyses of changes in community structure. *Australian Journal of Ecology* 18, 117–143

Dexter E., Rollwagen-Bollens G. & Bollens S.M. (2018). The trouble with stress: A flexible method for the evaluation of nonmetric multidimensional scaling. *Limnology and Oceanography: Methods* 16, 434–443.

Limitations of NMDS

- Results dependent on initial and random configurations
- 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 (e.g. to yield lower STRESS1 value)
- Fit of environmental variables more difficult to interpret than for metric (unconstrained) methods

Despite the fact that the fit of environmental variables to ordered distances is more difficult to interpret for NMDS, the method was among the best-performing when used to explain the relationship of plants with elevation in a comparative analysis (Wildi 2018).

Wildi O. (2018). Evaluating the Predictive Power of Ordination Methods in Ecological Context. *Mathematics* 6, 295. https://doi.org/10.3390/math6120295

RDA, similarity measures, NMDS and multivariate GLMs

Contents

1. Constrained ordination and RDA
2. RDA assumptions and PRC
3. Similarity measures
4. Nonmetric multidimensional scaling (NMDS)
5. **Model-based ordination, Multivariate GLMs**

42

Dissimilarity- and algorithm-based ordination

Y = Response (sites x species) transformed or non-transformed

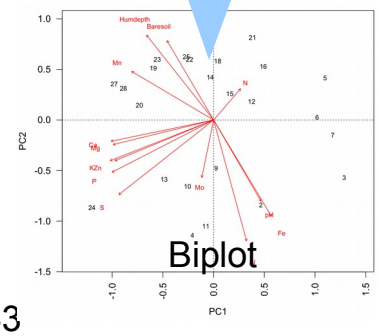
X = Predictors (sites x environmental vars.)

Unconstrained ordination (e.g. PCA, NMDS)

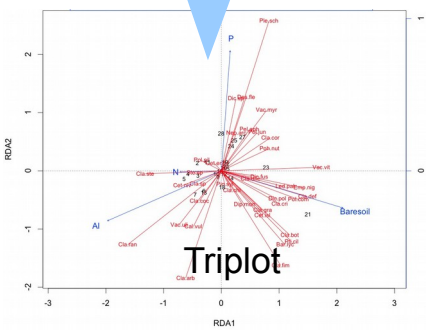
Constrained ordination (e.g. db-RDA, CCA)

$Ax = \lambda x$ Matrix decomposition

Algorithms on dissimilarity matrix



Biplot



Triplot

- Dominant approach in last decades
- Computationally very efficient
- Primarily applicable to research goals of explanation, exploration and assessing hypotheses and determining probabilities

43

In the literature, different terms are used to refer to a similar family of methods: dissimilarity-based (e.g. Anderson et al. 2019), distance-based (e.g. Roberts 2019, Warton et al. 2011) and algorithm-based (Warton et al. 2015) methods for ordination and multivariate analysis. We use the term dissimilarity-based and algorithm-based ordination here. Dissimilarity-based is used mainly for consistency with the chapter of similarity measures, following Anderson et al. (2019). Algorithm-based is used, following Warton et al. (2015), to highlight that these methods are rather defined by an algorithm than by a probabilistic framework (e.g. data model, parameter estimation and model selection based on maximum likelihood estimation).

Very loosely speaking the distinction between dissimilarity-based and algorithm-based methods on the one hand and model-based methods, which will be introduced in the following, on the other hand, follows a similar line of distinction as between algorithm-based and model-based methods made by Breimans classical paper on the two cultures of statistical modeling (Breiman 2001).

Importantly, methods such as PCA and CCA can be expressed in a probabilistic framework, hence their classification as algorithm-based rather refers to their classical use. For model-based versions of these methods see probabilistic PCA (Tipping & Bishop 1999) and constrained quadratic ordination (Yee 2004, Yee 2006).

The lack of a formal data model complicates prediction based on these type of methods.

Anderson M.J., de Valpine P., Punnett A. & Miller A.E. (2019). A pathway for multivariate analysis of ecological communities using copulas. *Ecology and Evolution* 9, 3276–3294. Freely accessible at: https://doi.org/10.1002/ece3.4948

Breiman L. (2001). Statistical modeling: The two cultures. *Statistical Science* 16, 199–215

Roberts D.W. (2019). Comparison of distance-based and model-based ordinations. *Ecology*, in press. Freely accessible within our university at: https://esajournals.onlinelibrary.wiley.com/doi/full/10.1002/ecy.2908

Tipping M.E. & Bishop C.M. (1999). Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 61, 611–622. Freely accessible at: http://www.robots.ox.ac.uk/~cvrg/hilary2006/ppca.pdf

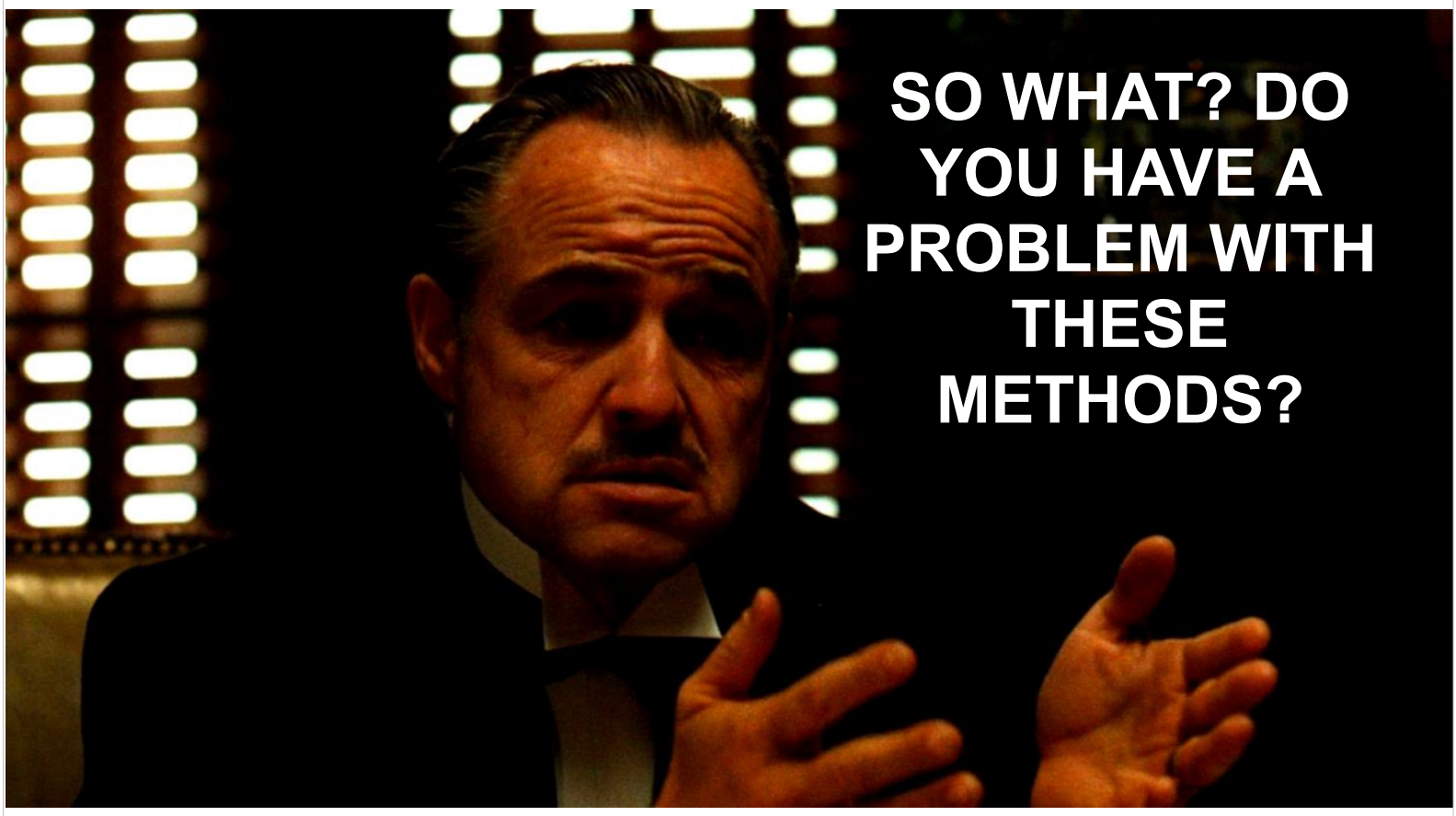
Warton D.I., Wright S.T. & Wang Y. (2011). Distance-based multivariate analyses confound location and dispersion effects. *Methods in Ecology and Evolution* 3, 89–101. Freely accessible at: https://doi.org/10.1111/j.2041-210X.2011.00127.x

Warton D.I., Foster S.D., De'ath G., Stoklosa J. & Dunstan P.K. (2015). Model-based thinking for community ecology. *Plant Ecology* 216, 669–682. Freely accessible within our university at: https://doi.org/10.1007/s11258-014-0366-3

Yee T.W. (2004). A new technique for maximum-likelihood canonical Gaussian ordination. *Ecological Monographs* 74, 685–701. Freely accessible within our university at: https://doi.org/10.1890/03-0078

Yee T.W. (2006). Constrained additive ordination. *Ecology* 87, 203–213. Freely accessible within our university at: https://doi.org/10.1890/05-0283

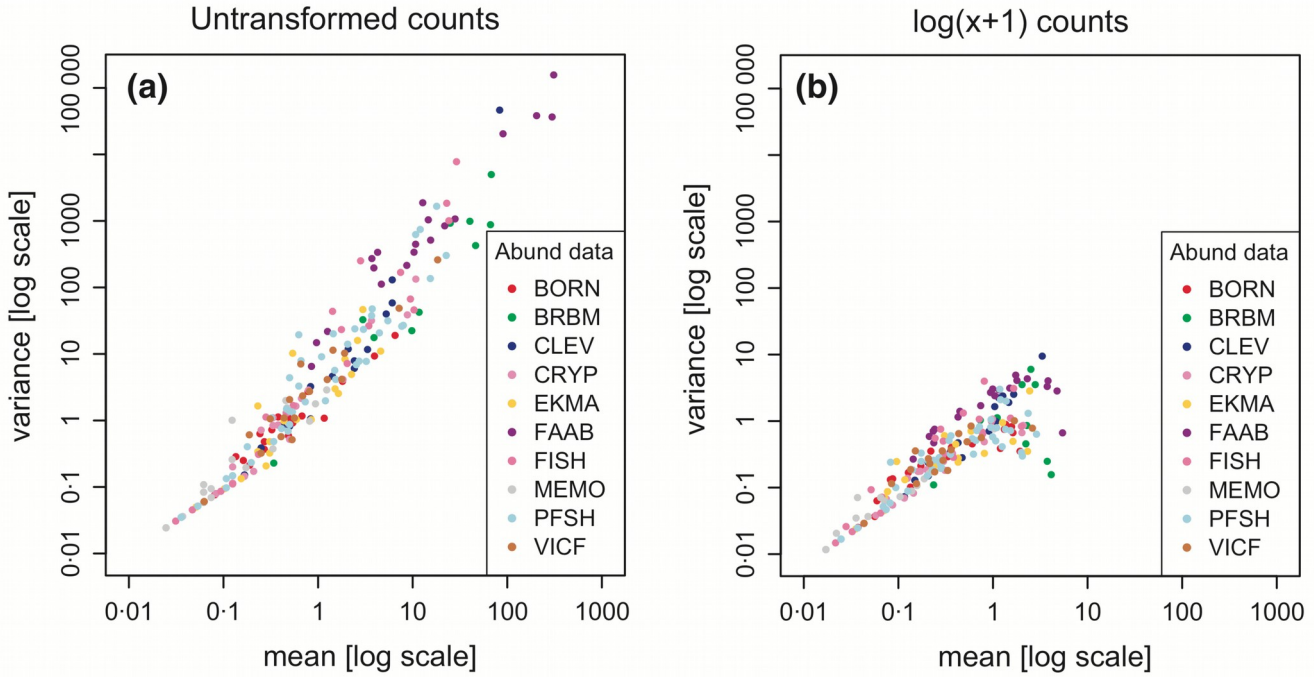
Dissimilarity- and algorithm-based ordination



44

Criticism of dissimilarity- and algorithm-based ordination

- Lack of probabilistic framework (e.g. statistical data model)
- Less interpretable due to lack of parameters at level of observations
- Ignorance of mean-variance relationship



Warton et al. 2011 *Meth. Ecol. Evol.* 3: 89

Most dissimilarity-based and algorithm-based methods lack a strictly defined statistical data model with associated model diagnosis, i.e. diagnosis to which extent the data are matching the assumptions. Statistical data models have been thoroughly studied and their assumptions are well known, whereas this is not the case for several of the similarity measures (Hui et al. 2014, Warton & Hui 2017). Given the lack of a data model, dissimilarity-based and algorithm-based approaches, in particular in the unconstrained case, do not distinguish between signal and noise (Hui et al. 2014). Moreover, the best-fit models are frequently found based on simulations (e.g. permutations) and not based on a classical probabilistic framework for model selection that relies, for instance, on maximum likelihood (e.g. AIC, BIC) and provides a statistic (e.g. deviance in case of GLM). For example, Hui et al. (2014) describe model-based approaches for unconstrained ordination where the number of dimensions can be decided based on the BIC, whereas this model selection criterion is not directly applicable to PCA (but to probabilistic PCA) and NMDS.

Warton et al. (2015) argue that the dissimilarity-based and algorithm-based models are less well interpretable because they lack parameters (or other quantities) on the level of the observations (e.g. species in ecology). In addition, they suggest that linking dissimilarity-based and algorithm-based models to ecological theory would be more difficult given this lack of parameters. By contrast, Roberts (2020) argues that the two approaches represent different theoretical perspectives in ecology, where one focuses on communities as a whole and the other on individual species, where communities are only an amalgamate of multiple species. However, the model-based approach can take both perspectives (Hui et al. 2014).

None of the methods that we refer to as dissimilarity-based and algorithm-based explicitly accounts for the mean-variance relationship, which is regarded as a major problem by Warton et al. (2011).

Hui F.K.C., Taskinen S., Pledger S., Foster S.D. & Warton D.I. (2014). Model-based approaches to unconstrained ordination. *Methods in Ecology and Evolution* 6, 399–411. reely accessible within our university at: https://doi.org/10.1111/2041-210X.12236

Roberts D.W. (2020). Comparison of distance-based and model-based ordinations. *Ecology* in press. Freely accessible within our university at: https://doi.org/10.1002/ecy.2908

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. Freely accessible at: https://doi.org/10.1111/j.2041-210X.2011.00127.x

Warton D.I., Foster S.D., De'ath G., Stoklosa J. & Dunstan P.K. (2015). Model-based thinking for community ecology. *Plant Ecology* 216, 669–682. Freely accessible within our university at: https://doi.org/10.1007/s11258-014-0366-3

Warton D.I. & Hui F.K.C. (2017). The central role of mean-variance relationships in the analysis of multivariate abundance data: a response to RoberWats (2017). *Methods in Ecology and Evolution* 8, 1408–1414. Freely accessible within our university at: https://doi.org/10.1111/2041-210X.12843

Model-based ordination

Y = Response (sites x species) transformed or non-transformed

X = Predictors (sites x environmental vars.)

Unconstrained analysis (e.g. GLLVMs, UAO)

Constrained analysis (e.g. multivariate GLMs, CAO)

- Model fitting for each response
  $g(\mu) = \beta_0 + ...$
- Model diagnosis
  $\mathrm{Var}(Y) = \phi\, V(\mu)$
- Inference based on individual responses
  Sum of LR-statistic, $\mathbf{B}_{(-1)} = \mathbf{C}\mathbf{A}^{T}$

- For several models: Ordination biplot
- Parameter estimates with confidence intervals
- Often likelihood statistics for model

- Emerging approach
- Computationally heavy (not feasible in the last century)
- Applicable to a wide range of research goals
- Explicitly defines signal (systematic model component) and noise (assumed distribution of error)

46

GLLVMs = Generalized linear latent variable models (e.g. Niku et al. 2019)
UAO = Unconstrained additive ordination (e.g. Yee 2006)

Most methods fit a GLM to each response, where $g()$ stands for the link function. The "..." in the equation stands for additional terms that the different methods imply (e.g. "$+ \beta_i x_i + \lambda z_i$" to add measured predictors and unmeasured predictors (also called latent variable) in the case of GLLVMs, "$f_i(x_i)$" to add a smoothing function for a (measured) predictor in the case of CAO). Model diagnosis also differs between methods, but would in most cases include checking whether the assumed mean-variance relationship holds. Inference is based on individual responses. For example, they are subject to matrix decomposition (e.g. Constrained linear, quadratic or additive ordination (Yee & Hastie 2003, Yee 2006)) or the test statistics for the individual responses are summed up (e.g. in the context of multivariate GLMs (Warton et al. 2011)). Note that RDA would also fall within the domain of model-based ordination, though as classically used and described in text books, model diagnosis is rarely done. CCA represents a heuristic method that simplifies constrained quadratic ordination (Yee 2004), which was to computationally heavy in the past.

Model-based ordination is an active field of research and rapidly growing. A very readable introduction is given in Warton et al. (2015) and for unconstrained analysis in Hui et al. (2014).

Hui F.K.C., Taskinen S., Pledger S., Foster S.D. & Warton D.I. (2014). Model-based approaches to unconstrained ordination. *Methods in Ecology and Evolution* 6, 399–411. Freely accessible within our university at: https://doi.org/10.1111/2041-210X.12236

Niku J., Hui F.K.C., Taskinen S. & Warton D.I. (2019). gllvm: Fast analysis of multivariate abundance data with generalized linear latent variable models in R. *Methods in Ecology and Evolution* 10, 2173-2182. Freely accessible within our university at: https://doi.org/10.1111/2041-210X.13303

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. Freely accessible at: https://doi.org/10.1111/j.2041-210X.2011.00127.x

Warton D.I., Foster S.D., De'ath G., Stoklosa J. & Dunstan P.K. (2015). Model-based thinking for community ecology. *Plant Ecology* 216, 669–682. Freely accessible within our university at: https://doi.org/10.1007/s11258-014-0366-3

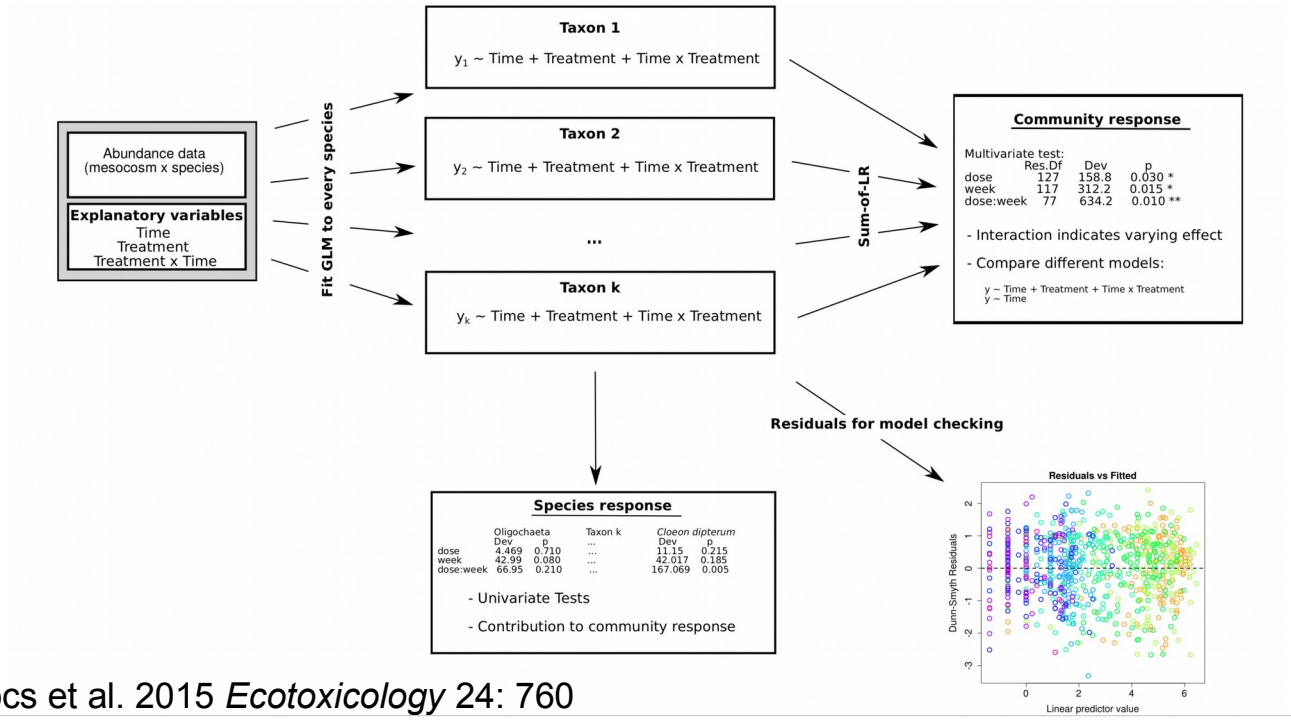
Yee T.W. & Hastie T.J. (2003). Reduced-rank vector generalized linear models. *Statistical Modelling* 3, 15–41. Freely accessible within our university at: https://doi.org/10.1191/1471082X03st045oa

Yee T.W. (2004). A new technique for maximum-likelihood canonical Gaussian ordination. *Ecological Monographs* 74, 685–701. Freely accessible within our university at: https://doi.org/10.1890/03-0078

Yee T.W. (2006). Constrained additive ordination. *Ecology* 87, 203–213. Freely accessible within our university at: https://doi.org/10.1890/05-0283

Multivariate GLMs (*sensu* Warton et al. 2011)

- Fit GLM for each response (note analogy to RDA)
- Diagnose model (e.g. check mean-variance assumption)
- No ordination biplot, mainly for assessing hypotheses, prediction and parameter estimation
- Can fit model analogous to PRC



47 Szöcs et al. 2015 *Ecotoxicology* 24: 760

We describe multivariate GLMs as introduced by Warton et al. (2011), see also Wang et al. (2012). A tutorial for the application and a comparison to PRCs is given in Szöcs et al. (2015).

Note that the sum of likelihood ratio statistic does not account for correlation between species, but the related p-value does. Similarly, the classical model framework for model selection is available (e.g. AIC, BIC), but ignores correlation among species. A different test statistic can be used to account for correlation in the test statistic. This issue is discussed in more detail in the following.

Szöcs E. et al. (2015) Analysing chemical-induced changes in macroinvertebrate communities in aquatic mesocosm experiments: a comparison of methods. *Ecotoxicology* 24, 760–769. Freely accessible within our university at: http://dx.doi.org/10.1007/s10646-015-1421-0

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. Freely available at: https://doi.org/10.1111/j.2041-210X.2012.00190.x

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. Freely accessible at: https://doi.org/10.1111/j.2041-210X.2011.00127.x

Multivariate GLMs and more complex models

How to account for the correlation between responses (e.g. species) in multivariate GLMs and other models?

- Estimate correlation from data and consider in model → Sample Variance-Covariance matrix $\boldsymbol{S}$ (unless true matrix $\boldsymbol{\Sigma}$ known). Only reliable if few responses or n >> number of responses.
- Assume no correlation (i.e. $\boldsymbol{\Sigma} = \boldsymbol{I}$). In most cases assumption wrong, but reliable p-values can be obtained through design-based inference (based on permutation)
- More complex procedures: Dimension reduction of $\boldsymbol{\Sigma}$ via latent variables, Generalized estimation equations, Copulas.

The issue of sample size becomes clear when considering the number of elements that need to be estimated in the variance-covariance matrix $\boldsymbol{\Sigma}$. In the case of 5 response variables (e.g. species), we would have 5 variances and 4 + 3 + 2 + 1 covariances for the pairs of response variables. We can calculate the number of elements in $\boldsymbol{\Sigma}$ for p responses as follows: $\frac{(p-1)\,p}{2}+p$

This means that the number of parameters grows quadratically. For example, for 20 species, which is not unusual (and rather a low number for invertebrates), we would need to estimate 210 parameters! It should be obvious that this is problematic if you have just tens of observations.

For details on the options to deal with correlation in multivariate GLMs see Wang et al. (2012). Dimension reduction via latent variables is demonstrated in Warton et al. (2015). General estimation equations are described in Warton et al. (2011).

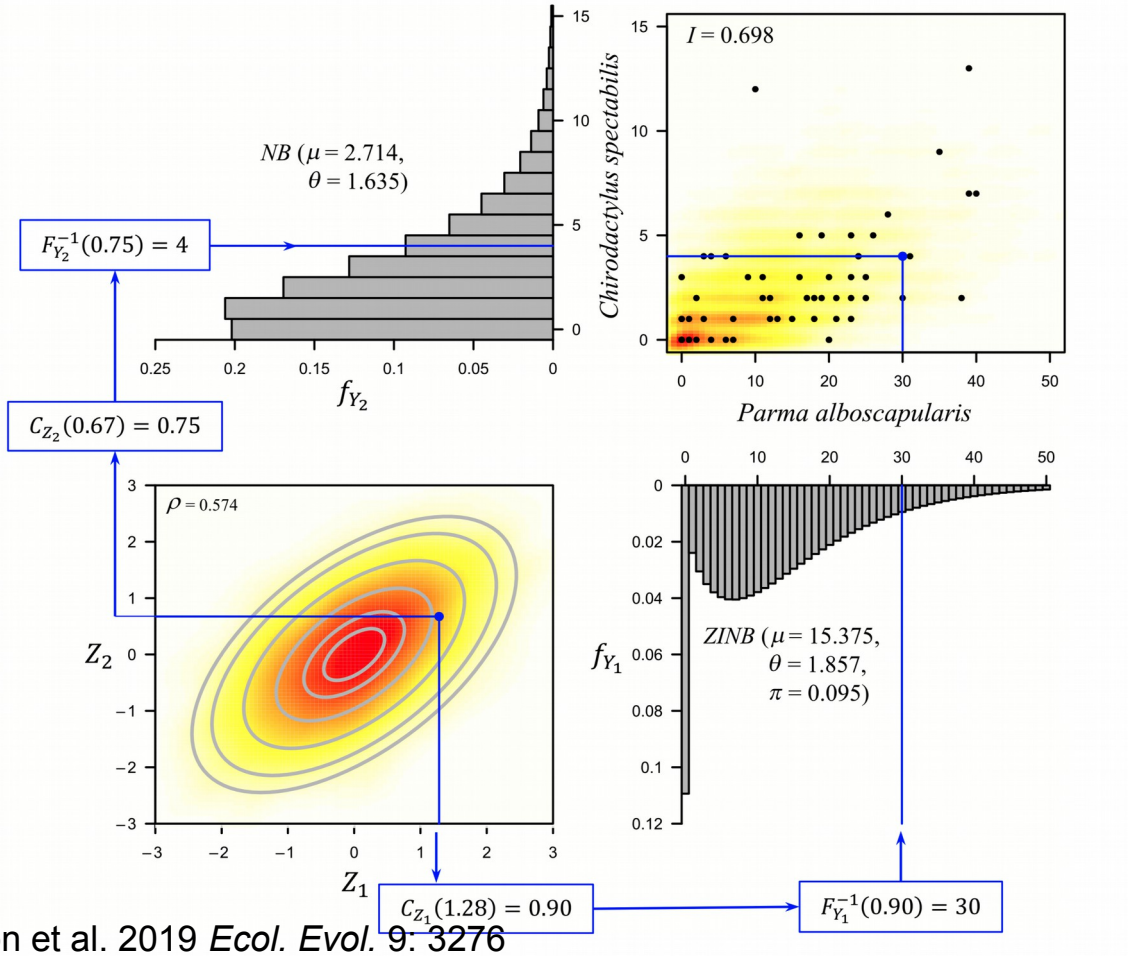
Wang Y., Naumann U., Wright S.T. & Warton D.I. (2012). mvabund – an R package for model-based analysis of multivariate abundance data. *Methods in Ecology and Evolution* 3, 471–474. Freely available at: https://doi.org/10.1111/j.2041-210X.2012.00190.x

Warton D.I. (2011). Regularized Sandwich Estimators for Analysis of High-Dimensional Data Using Generalized Estimating Equations. Biometrics 67, 116–123. https://doi.org/10.1111/J.1541-0420.2010.01438.X

Warton D.I., Blanchet F.G., O'Hara R.B., Ovaskainen O., Taskinen S., Walker S.C., et al. (2015). So Many Variables: Joint Modeling in Community Ecology. Trends in Ecology & Evolution 30, 766–779. https://doi.org/10.1016/j.tree.2015.09.007

Copulas for modelling species interactions

Copula: Function representing joint distribution based on the cumulative distribution functions of individual variables → can couple variables



49 Anderson et al. 2019 *Ecol. Evol.* 9: 3276

The two coefficients in the figure are the Pearson correlation coefficient ρ and the inter-specific index of association I.

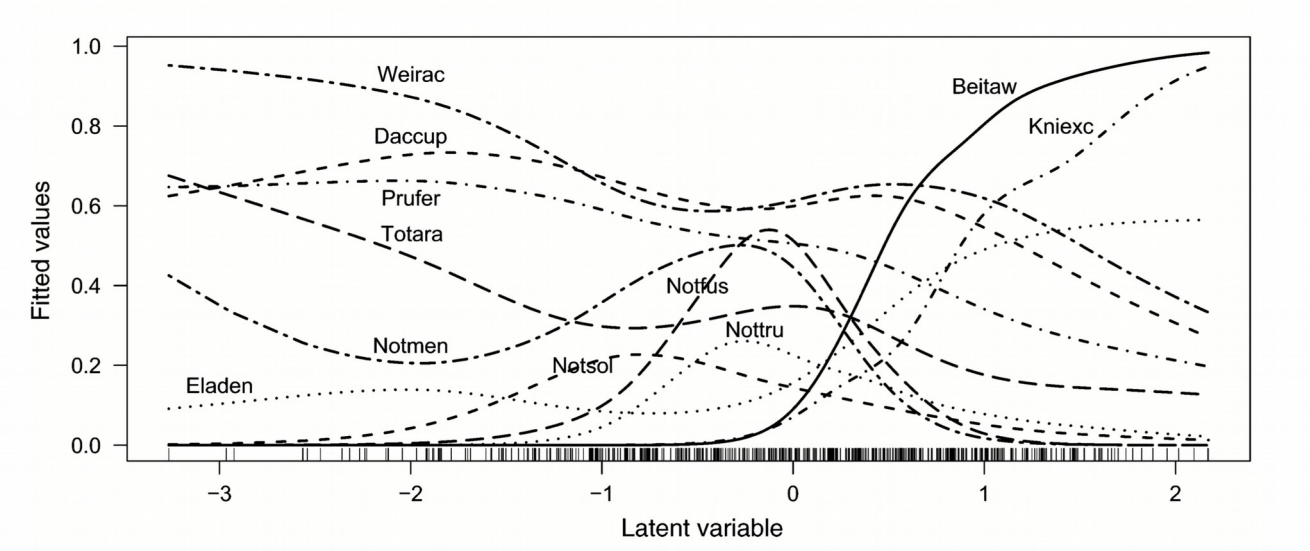
Anderson et al. 2019 use copulas in concert with a dissimilarity-based method (i.e. MDS, Permutational Multivariate ANOVA (PERMANOVA)). Popovic et al. (2019) also use copulas but in a purely model-based context for Gaussian copula graphical models (GCGMs).

Anderson M.J., de Valpine P., Punnett A. & Miller A.E. (2019). A pathway for multivariate analysis of ecological communities using copulas. *Ecology and Evolution* 9, 3276–3294. Freely accessible at: https://doi.org/10.1002/ece3.4948

Popovic G.C., Warton D.I., Thomson F.J., Hui F.K.C. & Moles A.T. (2019). Untangling direct species associations from indirect mediator species effects with graphical models. *Methods in Ecology and Evolution* 10, 1571–1583. Freely accessible within our university at: https://doi.org/10.1111/2041-210X.13247

Constrained Additive ordination

- Data-based rather than model-based ordination: Relies on Generalized Additive Models $g(\mu) = \beta_0 + f(v)$
- Derives response of each species to main environmental gradient from data → no linear or unimodal model assumed, currently restricted to one latent variable
- Computationally very demanding
- Example: Several species deviate from linear and unimodal shape



50 Yee 2006 *Ecology* 87: 203

CAO represents a novel and flexible ordination approach (Yee 2006, Yee 2015) that can be used for any species response to environmental gradients. However, it is currently limited to one latent variable that is extracted from the set of predictors. 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 the main gradient. The method is computationally demanding (calculations can take up to several hours, depending on the model specification and size of the data set). For details on the implementation in R with example code refer to Yee (2015).

Yee T.W. (2006). Constrained additive ordination. *Ecology* 87, 203–213. https://doi.org/10.1890/05-0283

Yee T.W. (2015). Vector generalized linear and additive models: with an implementation in R. Springer, New York, NY.

Goodbye, dissimilarity-based methods?

- Comparisons between methods paint a more nuanced picture, model-based approaches not always superior
- Joint approaches (e.g. Anderson et al. 2019)
- Mind your data properties (e.g. mean-variance relationship, response shape) and research goal/question!

https://www.needpix.com/photo/download/935292/puppy-dogs-collie-cute-pet-sweet-free-pictures-free-photos-free-images

A few studies compared the performance of dissimilarity- and algorithm-based methods and model-based methods for simulated and real world data sets. Jupke and Schäfer (2020) found that both multivariate GLMs and distance-based RDA reliably discriminated noise from true environmental variables, whereas Constrained quadratic ordination (CQO) and CCA performed worse. Similarly, Szöcs et al. (2015) found a similar performance of multivariate GLMs and PRC when used to analyse data from ecotoxicological experiments. Roberts (2020) concluded that two dissimilarity-based methods (including NMDS) outperformed two model-based methods (both not discussed here: BORAL, which is a Bayesian version of latent variable ordination and analysis, and REO, which relates to CQO). Warton et al. (2011) demonstrated that multivariate GLMs were more reliable than dissimilarity-based methods, RDA and CCA for the analysis of selected data sets. Finally, Yamaura et al. (2019) found a similar performance of a model-based method and distance-based RDA, conditional on the distance measure. Hence, model-based methods are not generally superior to dissimilarity-based ones.

In particular in the case of complex non-linear relationships between environmental variables and the responses (e.g. community), dissimilarity- and algorithm-based methods may be a good choice, given the limitations of CAO (e.g. only one latent variable).

Anderson M.J., de Valpine P., Punnett A. & Miller A.E. (2019). A pathway for multivariate analysis of ecological communities using copulas. *Ecology and Evolution* 9, 3276–3294. Freely accessible at: https://doi.org/10.1002/ece3.4948

Jupke & Schäfer (2020). Should ecologists prefer model- over distance-based multivariate methods? *Ecol. Evol.* In press.

Roberts D.W. (2020). Comparison of distance-based and model-based ordinations. *Ecology.* In press. Freely accessible within our university at: https://doi.org/10.1002/ecy.2908

Szöcs E. et al. (2015) Analysing chemical-induced changes in macroinvertebrate communities in aquatic mesocosm experiments: a comparison of methods. *Ecotoxicology* 24, 760–769. Freely accessible within our university at: http://dx.doi.org/10.1007/s10646-015-1421-0

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. Freely accessible at: https://doi.org/10.1111/j.2041-210X.2011.00127.x

Yamaura Y., Blanchet F.G. & Higa M. (2019). Analyzing community structure subject to incomplete sampling: hierarchical community model vs. canonical ordinations. *Ecology* 100, e02759. Freely accessible within our university at: https://doi.org/10.1002/ecy.2759