RDA, similarity measures and NMDS

Contents

1. **Learning targets, constrained ordination and RDA**
2. Diagnosis and assumptions of RDA, extensions and orientation on stats methods
3. Similarity and distance measures
4. Non-metric multidimensional scaling (NMDS)

Learning targets

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

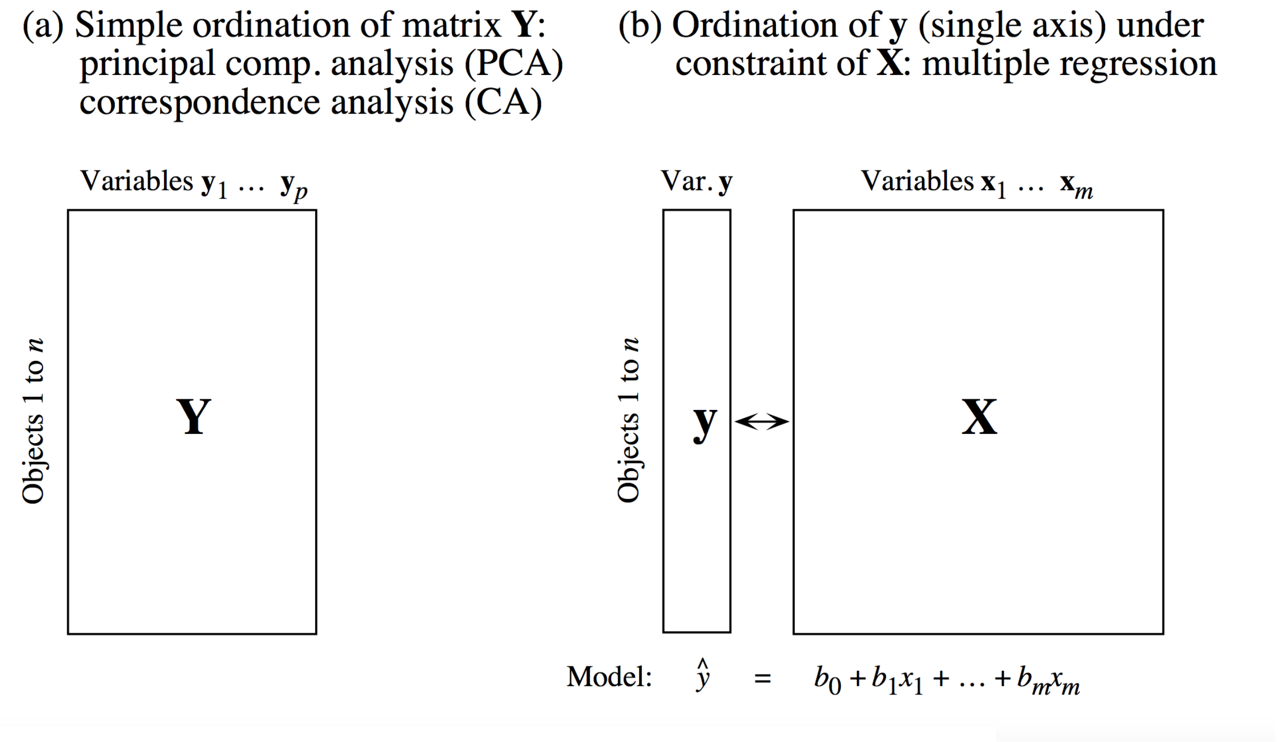
Learning targets and study questions

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

Learning targets and study questions

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

Constrained ordination methods



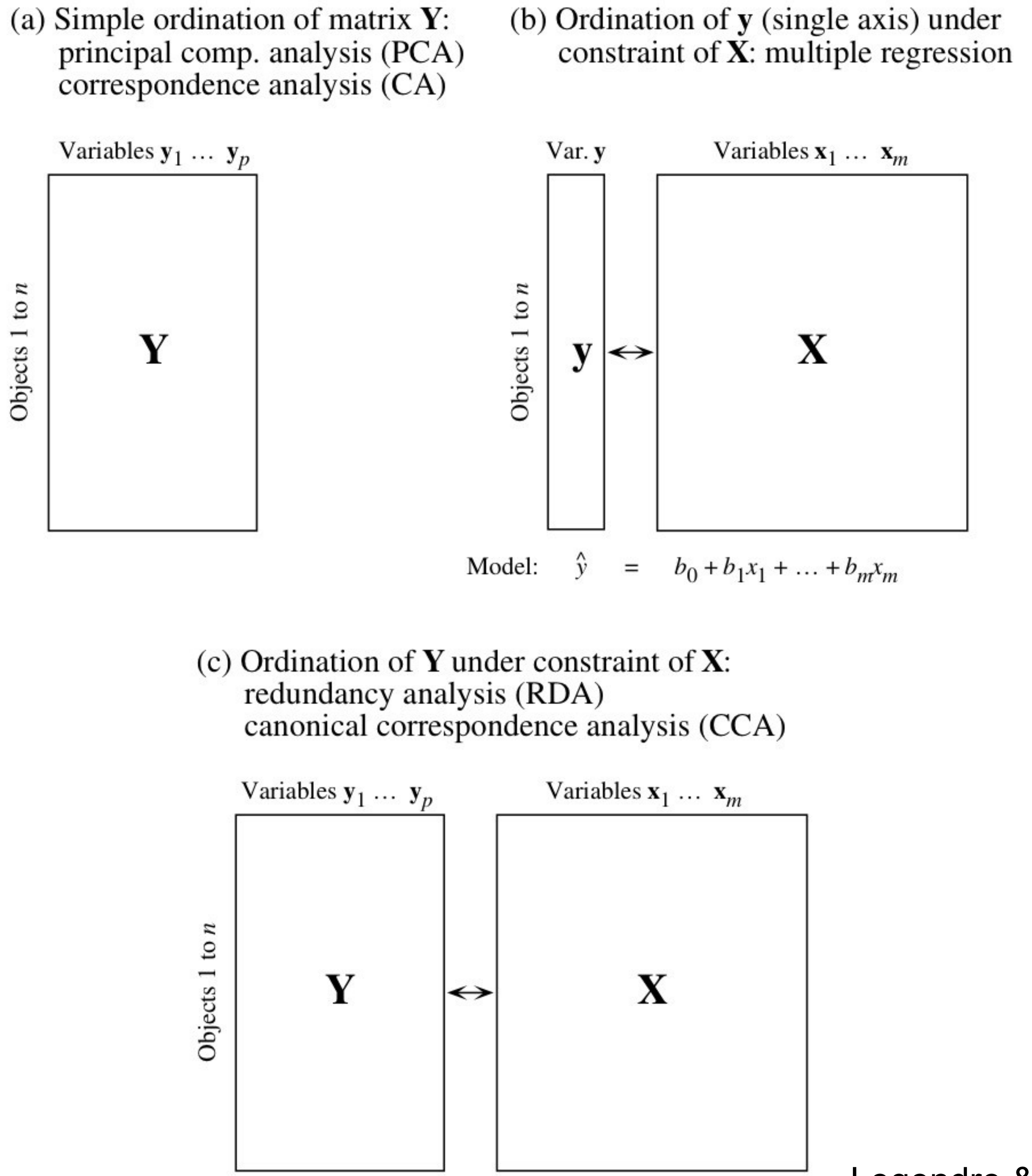
(a) Simple ordination of matrix **Y**: principal comp. analysis (PCA) correspondence analysis (CA)

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

Variables $\mathbf{y}_1 \dots \mathbf{y}_p$
Objects 1 to n
Y
Var. **y**
Variables $\mathbf{x}_1 \dots \mathbf{x}_m$
Objects 1 to n
y
X
Model: $\hat{y} = b_0 + b_1 x_1 + \dots + b_m x_m$

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

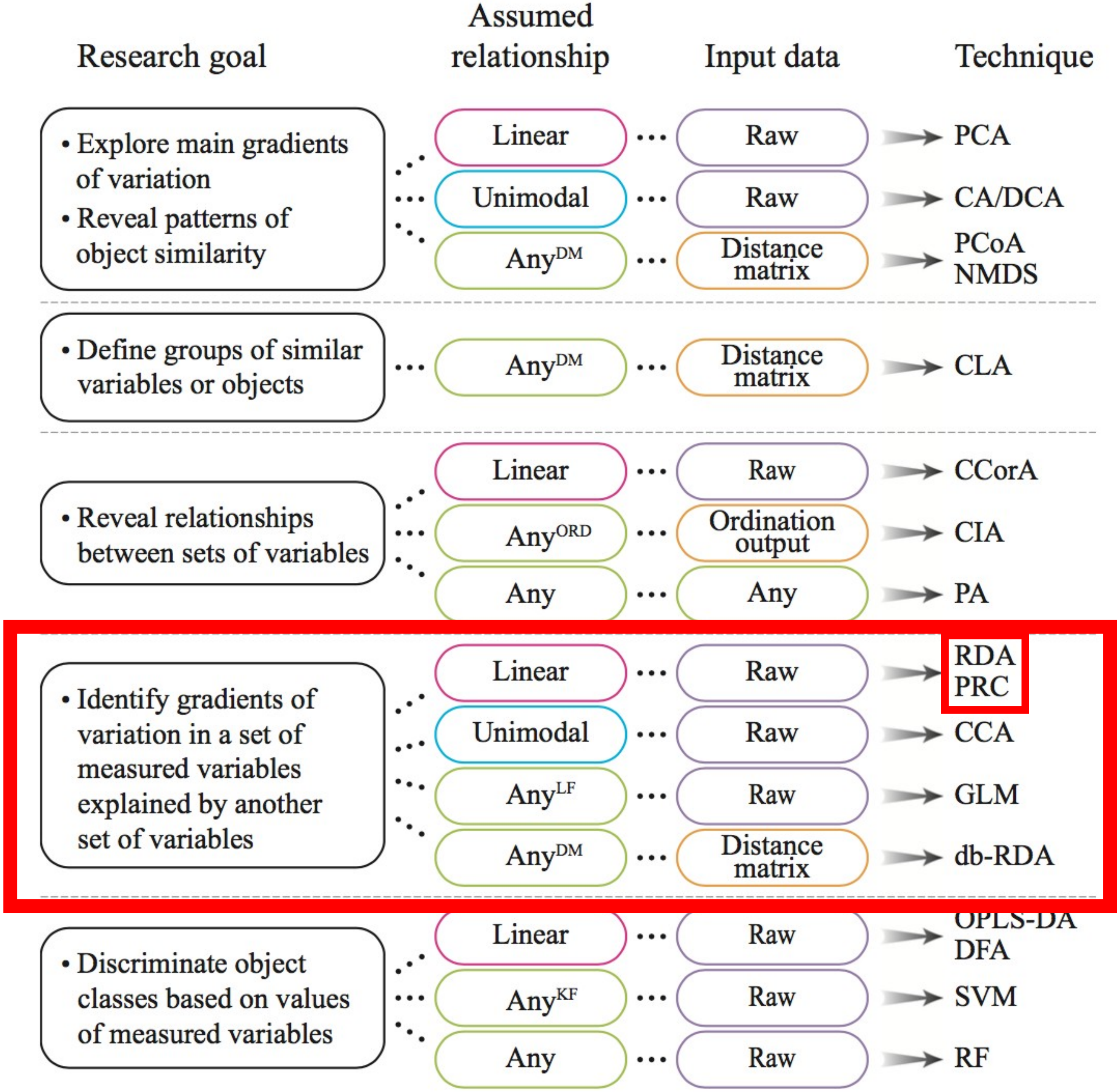
Model: $\hat{y} = b_0 + b_1x_1 + \ldots + b_mx_m$

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

Variables $\mathbf{y}_1 \ldots \mathbf{y}_p$
Objects 1 to n
Y
Var. **y**
Variables $\mathbf{x}_1 \ldots \mathbf{x}_m$
Objects 1 to n
y
X
Variables $\mathbf{y}_1 \ldots \mathbf{y}_p$
Variables $\mathbf{x}_1 \ldots \mathbf{x}_m$
Objects 1 to n
Y
X

Legendre & Legendre 2012: 627

Constrained ordination methods



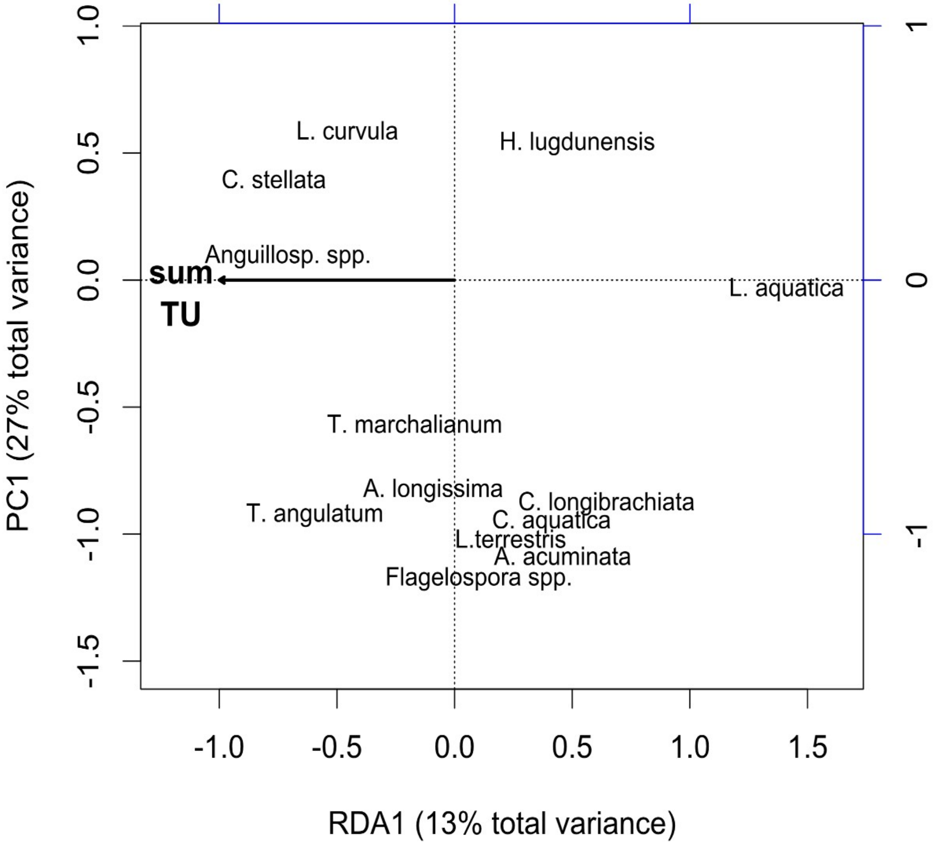
Research goal	Assumed relationship	Input data	Technique
• Explore main gradients of variation • Reveal patterns of object similarity	Linear	Raw	PCA
	Unimodal	Raw	CA/DCA
	Any^{DM}	Distance matrix	PCoA NMDS
• Define groups of similar variables or objects	Any^{DM}	Distance matrix	CLA
• Reveal relationships between sets of variables	Linear	Raw	CCorA
	Any^{ORD}	Ordination output	CIA
	Any	Any	PA
• Identify gradients of variation in a set of measured variables explained by another set of variables	Linear	Raw	RDA PRC
	Unimodal	Raw	CCA
	Any^{LF}	Raw	GLM
	Any^{DM}	Distance matrix	db-RDA
• Discriminate object classes based on values of measured variables	Linear	Raw	OPLS-DA DFA
	Any^{KF}	Raw	SVM
	Any	Raw	RF

Paliy & Shankar 2016 *Mol Ecol*:1032

Redundancy Analysis (RDA)

<u>Aim:</u> 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

$$\begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1\mathrm{m}} \\ 1 & x_{21} & x_{22} & \cdots & x_{2\mathrm{m}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_m \end{pmatrix}$$

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

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

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

Reformulation for the case of multivariate multiple regression with several ***y***:

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

Mathematical background of RDA

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

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

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

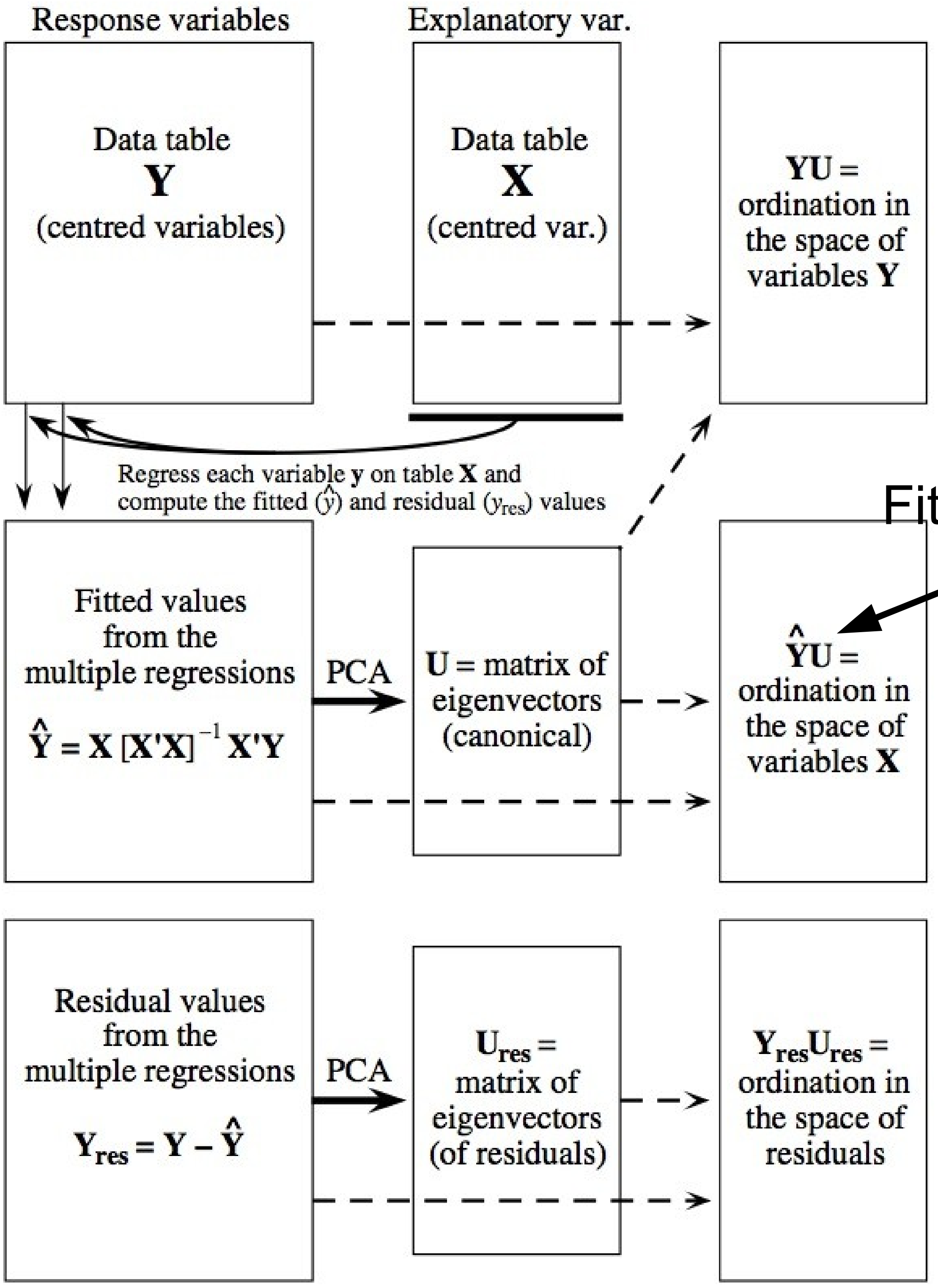
$$S_{\hat{Y}^t \hat{Y}} = \frac{1}{n-1} \hat{Y}^t \hat{Y}$$

and used in a PCA: $$\boxed{S_{\hat{Y}^t \hat{Y}}\, a = \lambda\, a}$$

Eigenvector

Eigenvalue problem

→ Eigenvectors linear combinations of predictors



Response variables

Explanatory var.

Data table $\mathbf{Y}$ (centred variables)

Data table $\mathbf{X}$ (centred var.)

$\mathbf{YU}$ = ordination in the space of variables $\mathbf{Y}$

Regress each variable $\mathbf{y}$ on table $\mathbf{X}$ and compute the fitted ($\hat{y}$) and residual (y_{res}) values

Fitted values from the multiple regressions

$\hat{\mathbf{Y}} = \mathbf{X}\,[\mathbf{X'X}]^{-1}\,\mathbf{X'Y}$

PCA

$\mathbf{U}$ = matrix of eigenvectors (canonical)

$\hat{\mathbf{Y}}\mathbf{U}$ = ordination in the space of variables $\mathbf{X}$

Fitted site scores

Residual values from the multiple regressions

$\mathbf{Y}_{\mathbf{res}} = \mathbf{Y} - \hat{\mathbf{Y}}$

PCA

$\mathbf{U}_{\mathbf{res}}$ = matrix of eigenvectors (of residuals)

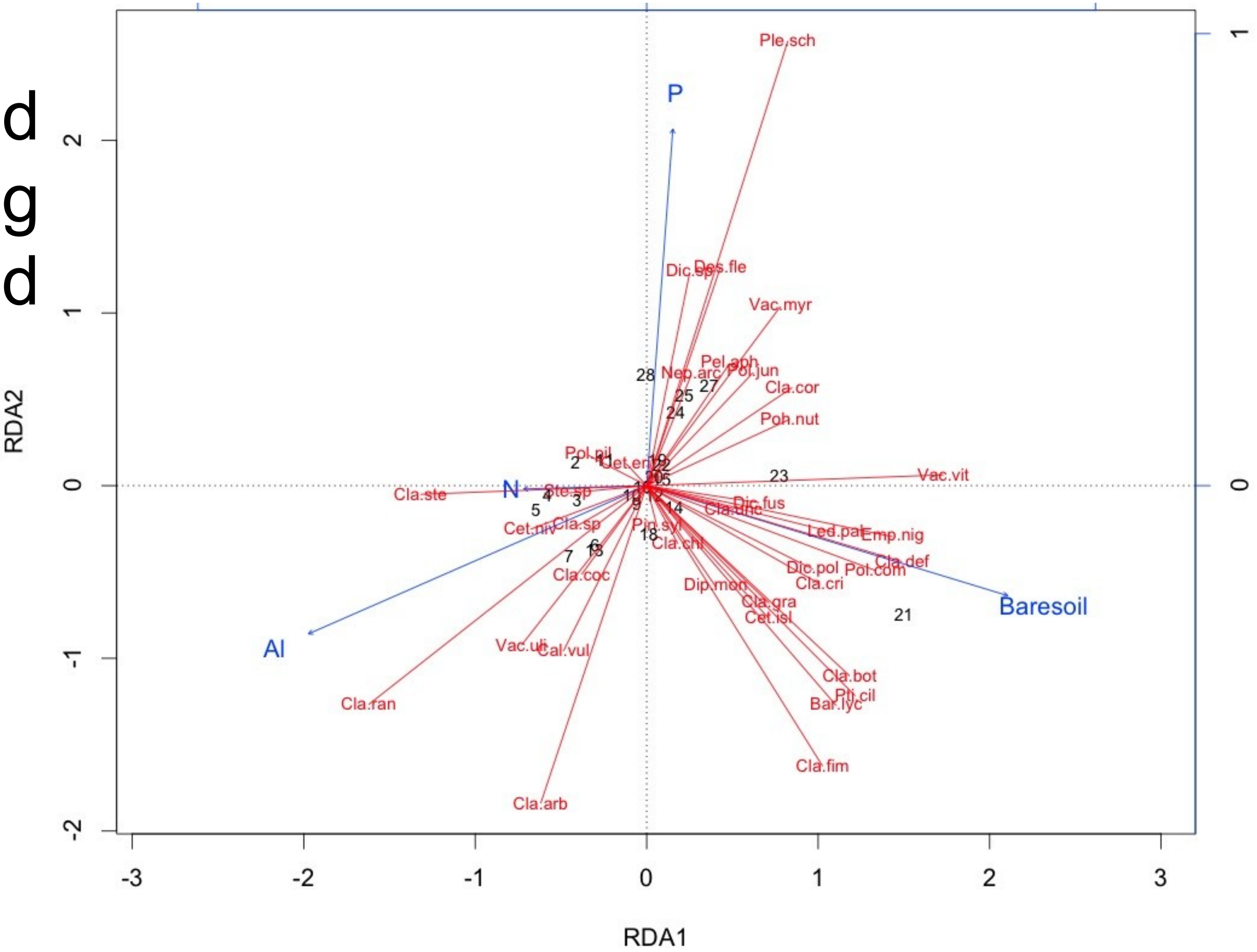
$\mathbf{Y}_{\mathbf{res}}\mathbf{U}_{\mathbf{res}}$ = ordination in the space of residuals

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 and NMDS

Contents

1. Learning targets, constrained ordination and RDA
2. **Diagnosis and assumptions of RDA, extensions and orientation on stats methods**
3. Similarity and distance measures
4. Non-metric multidimensional scaling (NMDS)

RDA axes and variable importance

How many RDA axes are required?

- Hypothesis test (permutation-based) recommended (Legendre et al. *MEE* 2011)

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

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

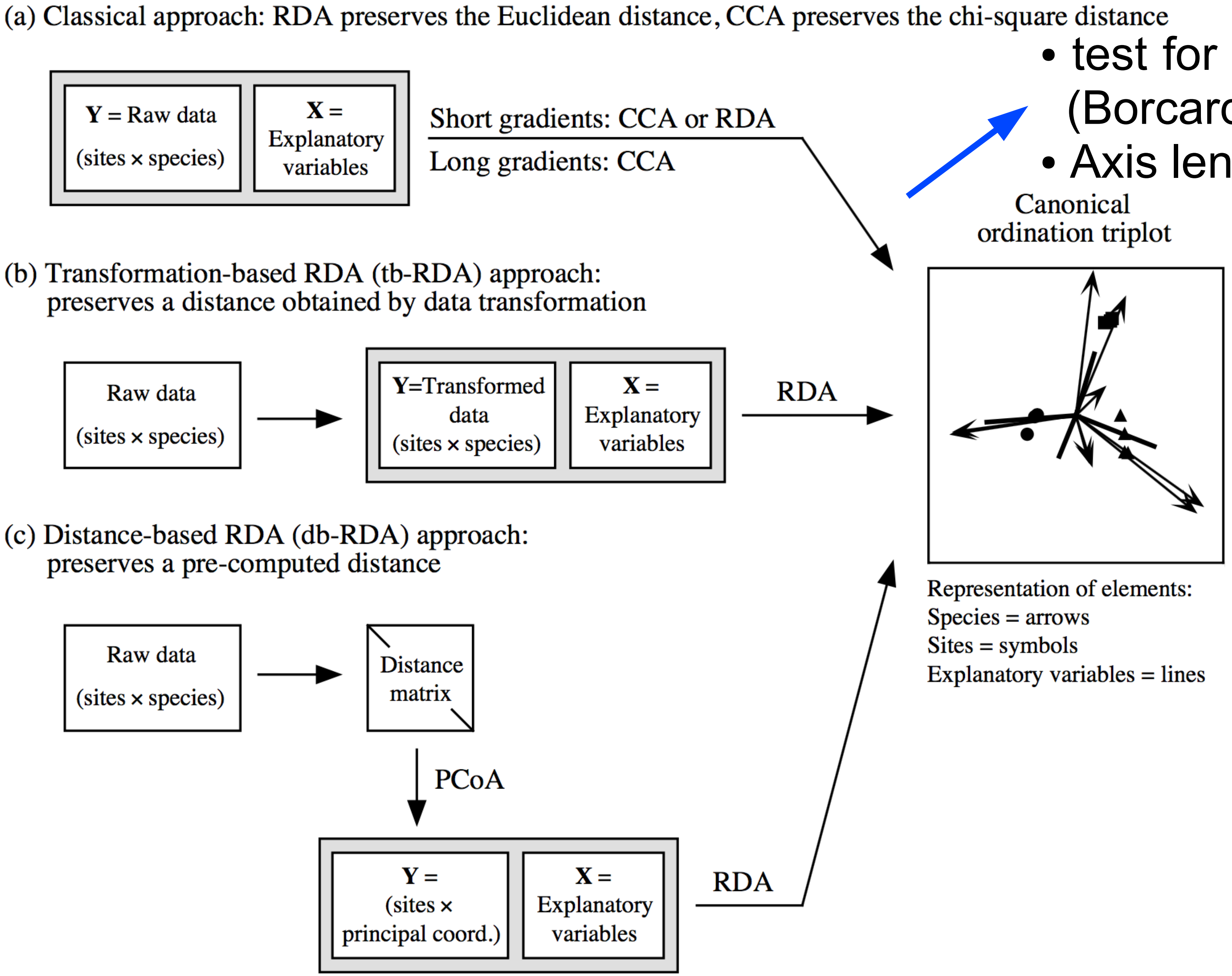
Assumptions and extensions of RDA

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

RDA approaches

How to assess gradient length?

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



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

Y = Raw data (sites × species)

X = Explanatory variables

Short gradients: CCA or RDA

Long gradients: CCA

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

Raw data (sites × species)

Y=Transformed data (sites × species)

X = Explanatory variables

RDA

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

Raw data (sites × species)

Distance matrix

PCoA

Y = (sites × principal coord.)

X = Explanatory variables

RDA

Canonical ordination triplot

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

Legendre & Legendre 2012: 648

Further constrained ordination methods

Canonical Correspondence Analysis (CCA)

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

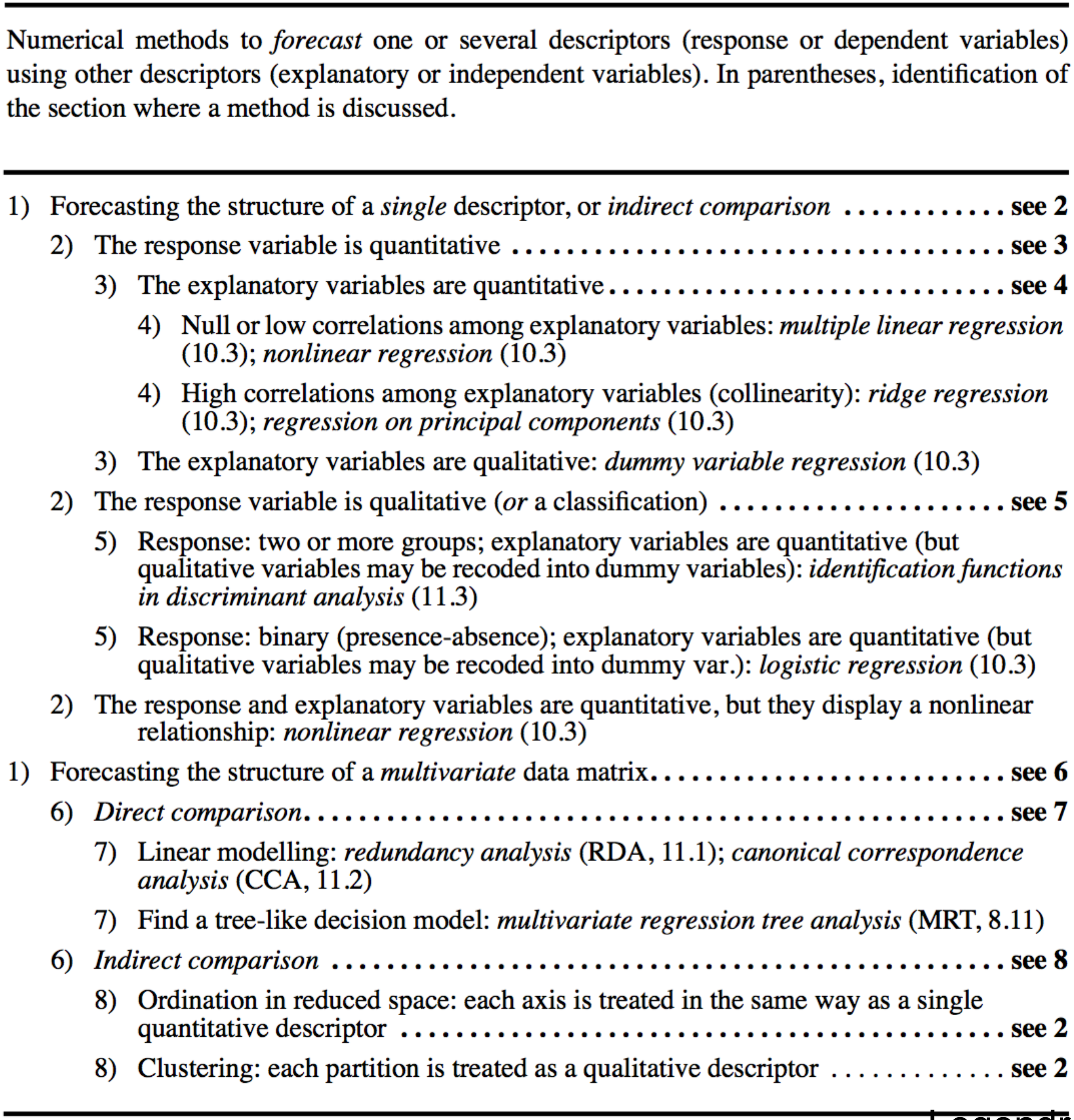
```
cca() {vegan}
```

Constrained additive Ordination (CAO)

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

```
cao() {VGAM}
```

When to use what?



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

1) Forecasting the structure of a *single* descriptor, or *indirect comparison* .... **see 2**

2) The response variable is quantitative .... **see 3**

3) The explanatory variables are quantitative.............................. **see 4**

4) Null or low correlations among explanatory variables: *multiple linear regression* (10.3); *nonlinear regression* (10.3)

4) High correlations among explanatory variables (collinearity): *ridge regression* (10.3); *regression on principal components* (10.3)

3) The explanatory variables are qualitative: *dummy variable regression* (10.3)

2) The response variable is qualitative (*or* a classification) .... **see 5**

5) Response: two or more groups; explanatory variables are quantitative (but qualitative variables may be recoded into dummy variables): *identification functions in discriminant analysis* (11.3)

5) Response: binary (presence-absence); explanatory variables are quantitative (but qualitative variables may be recoded into dummy var.): *logistic regression* (10.3)

2) The response and explanatory variables are quantitative, but they display a nonlinear relationship: *nonlinear regression* (10.3)

1) Forecasting the structure of a *multivariate* data matrix.......................... **see 6**

6) *Direct comparison*.. **see 7**

7) Linear modelling: *redundancy analysis* (RDA, 11.1); *canonical correspondence analysis* (CCA, 11.2)

7) Find a tree-like decision model: *multivariate regression tree analysis* (MRT, 8.11)

6) *Indirect comparison* .... **see 8**

8) Ordination in reduced space: each axis is treated in the same way as a single quantitative descriptor .... **see 2**

8) Clustering: each partition is treated as a qualitative descriptor .... **see 2**

Legendre & Legendre 2012: 533

RDA, similarity measures and NMDS

Contents

1. Learning targets, constrained ordination and RDA
2. Diagnosis and assumptions of RDA, extensions and orientation on stats methods
3. **Similarity and distance measures**
4. Non-metric multidimensional scaling (NMDS)

Measuring association

Example: Species observations in 4 streams

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

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

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

Similarity measures: Presence-Absence

Simple matching coefficient

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

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

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

Site				
1	0	400	0	0
2	0	0	10	0

Similarity measures: Presence-Absence

Site				
1	0	400	0	0
2	0	0	10	0

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

Calculation with all species:

a = 0, b = 1, c = 1, d = 2 → S_m = 2/4 = 0.5

Calculation without species 1 and 4:

a = 0, b = 1, c = 1, d = 0 → S_m = 0/2 = 0

Species absence influences similarity between sites.

Not desirable: joint absence of species does not indicate ecological similarity and number of joint absences is arbitrary → **Double-Zero problem**

Widely used similarity measures

Jaccard coefficient (=Jaccard similarity index)

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

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

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

Bray-Curtis coefficient

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

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

x_{ik} and x_{jk} is the abundance of taxon *k* for site *i and j*.

Example: Bray-Curtis coefficient

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

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

Calculation:

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

Calculation after square-root transformation:

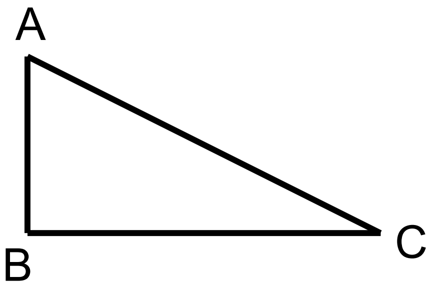
2*(0+0+$5^{0.5}$+0)/($400^{0.5}$+$5^{0.5}$+$10^{0.5}$) → S_{BC} = 0.18

Calculation after double square-root transformation:

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

Distance measures

Association measures meeting triangle inequality criterion
(following Everitt et al. 2011 *Cluster Analysis*. John Wiley & Sons: 49)



Triangle inequality criterion

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

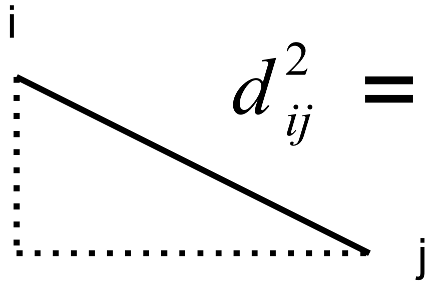
Sum of any two sides of triangle always ≥ third site

Important for geometrical representation (e.g. Ordination)

Euclidean distance: Most frequently used distance measure

$$d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}$$

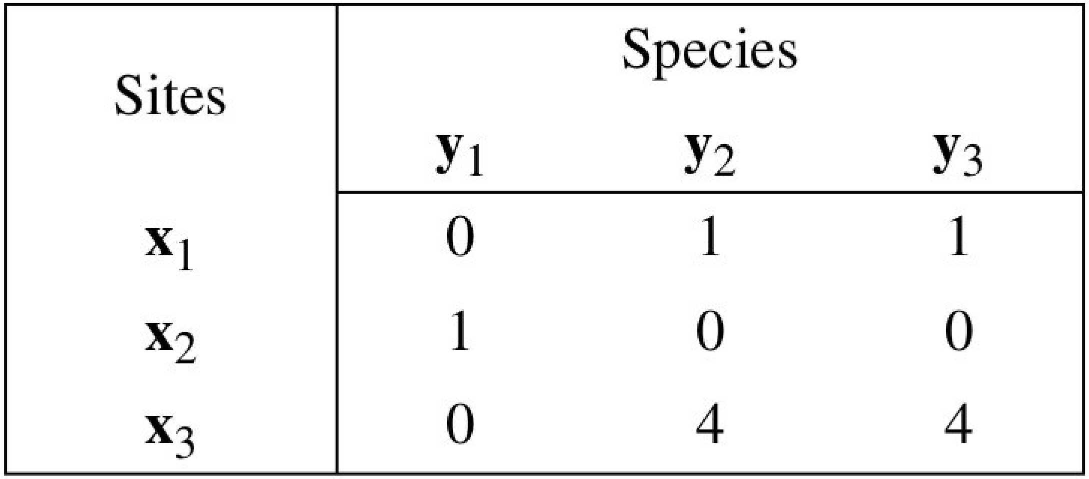
Two dimensional case:



$$d_{ij}^2 = (x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2$$

Species abundance paradox

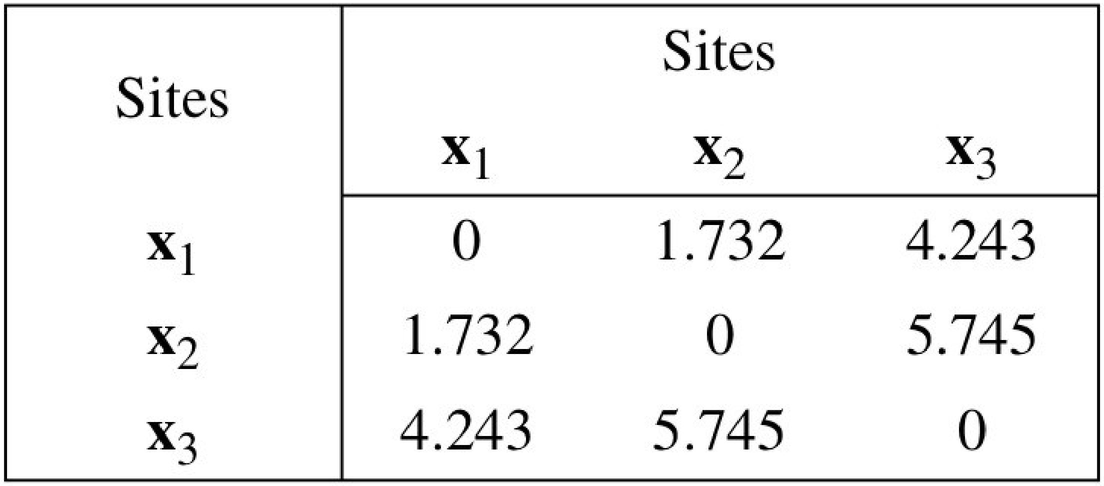
Species x Site matrix



Sites	Species		
	$\mathbf{y}_1$	$\mathbf{y}_2$	$\mathbf{y}_3$
$\mathbf{x}_1$	0	1	1
$\mathbf{x}_2$	1	0	0
$\mathbf{x}_3$	0	4	4

Euclidean distance

Distance matrix



Sites	Sites		
	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
$\mathbf{x}_1$	0	1.732	4.243
$\mathbf{x}_2$	1.732	0	5.745
$\mathbf{x}_3$	4.243	5.745	0

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

→ Euclidean distance problematic for ecological data

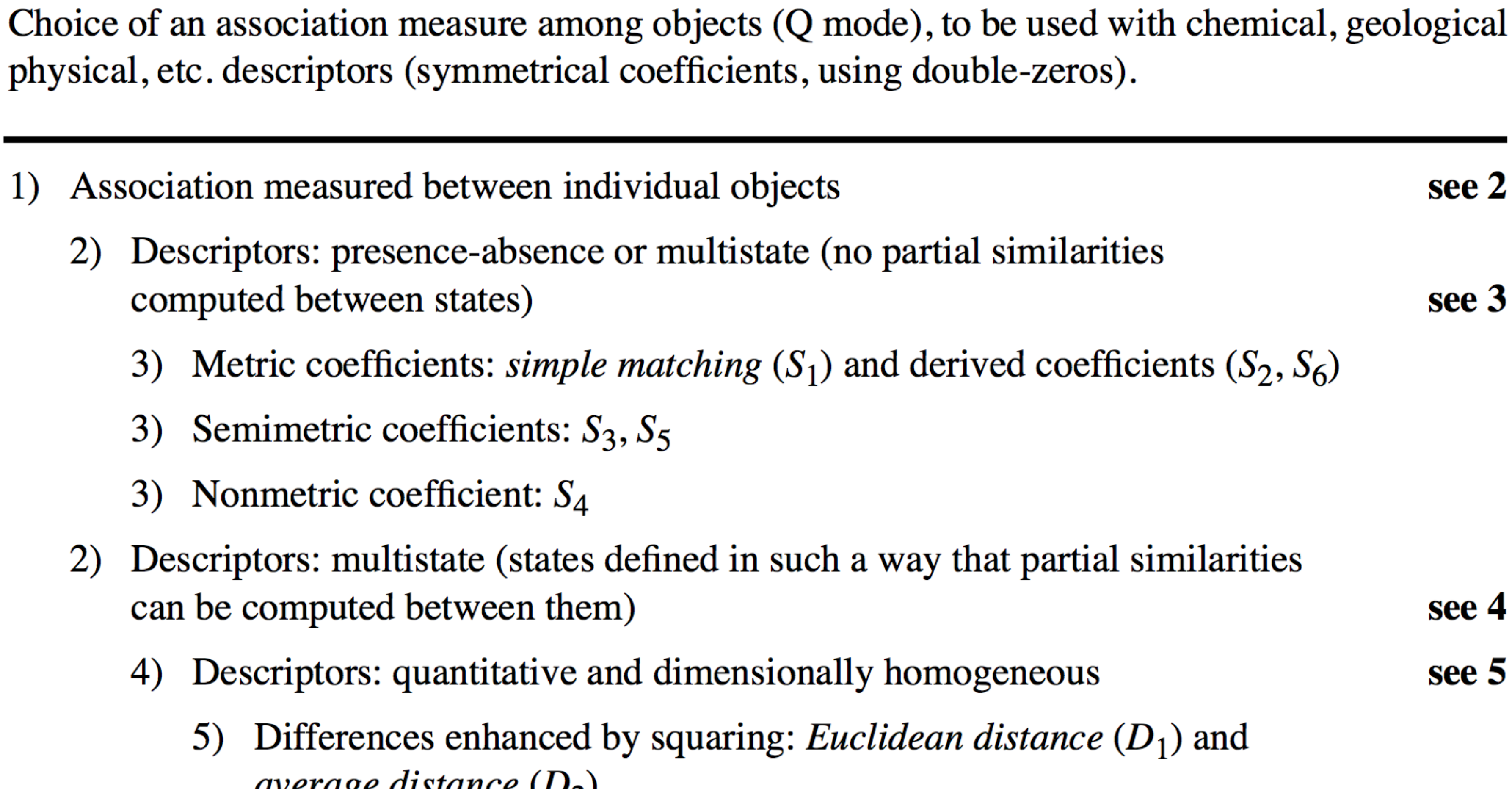
Legendre & Legendre 2012: 300

How to select a measure

- Many more association measures

(see Legendre & Legendre 2012: Chapter 7)

- check literature of scientific field
- 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)

Association measures in R

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

RDA, similarity measures and NMDS

Contents

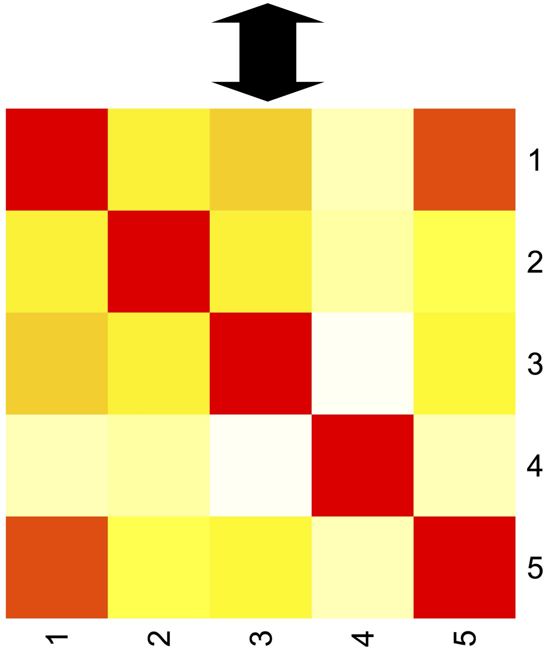
1. Learning targets, constrained ordination and RDA
2. Diagnosis and assumptions of RDA, extensions and orientation on stats methods
3. Similarity and distance measures
4. **Non-metric multidimensional scaling (NMDS)**

Visualization of association measures

Heatmap

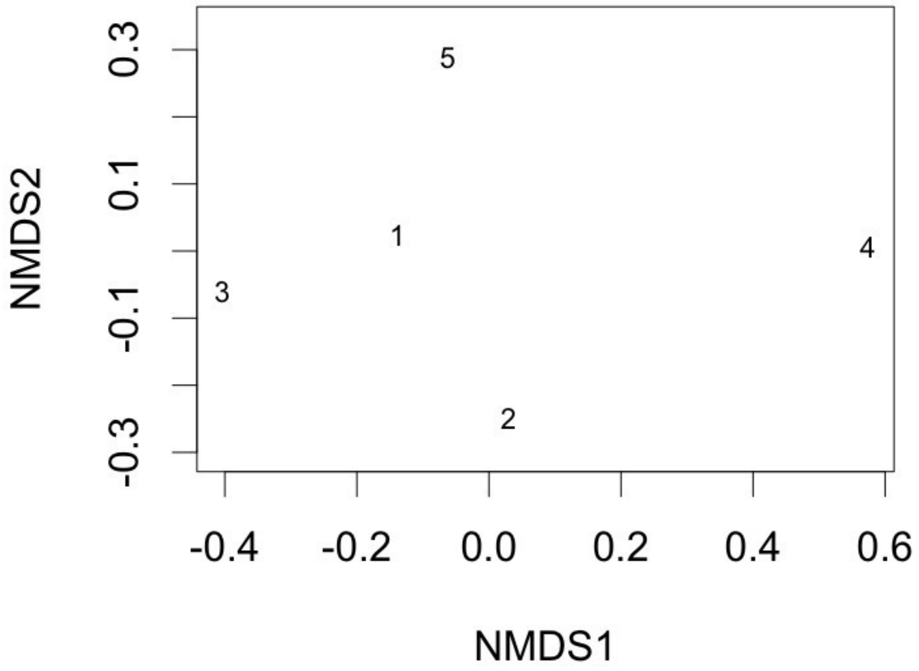
- Associations converted to colours
- Relationship easier to grasp

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

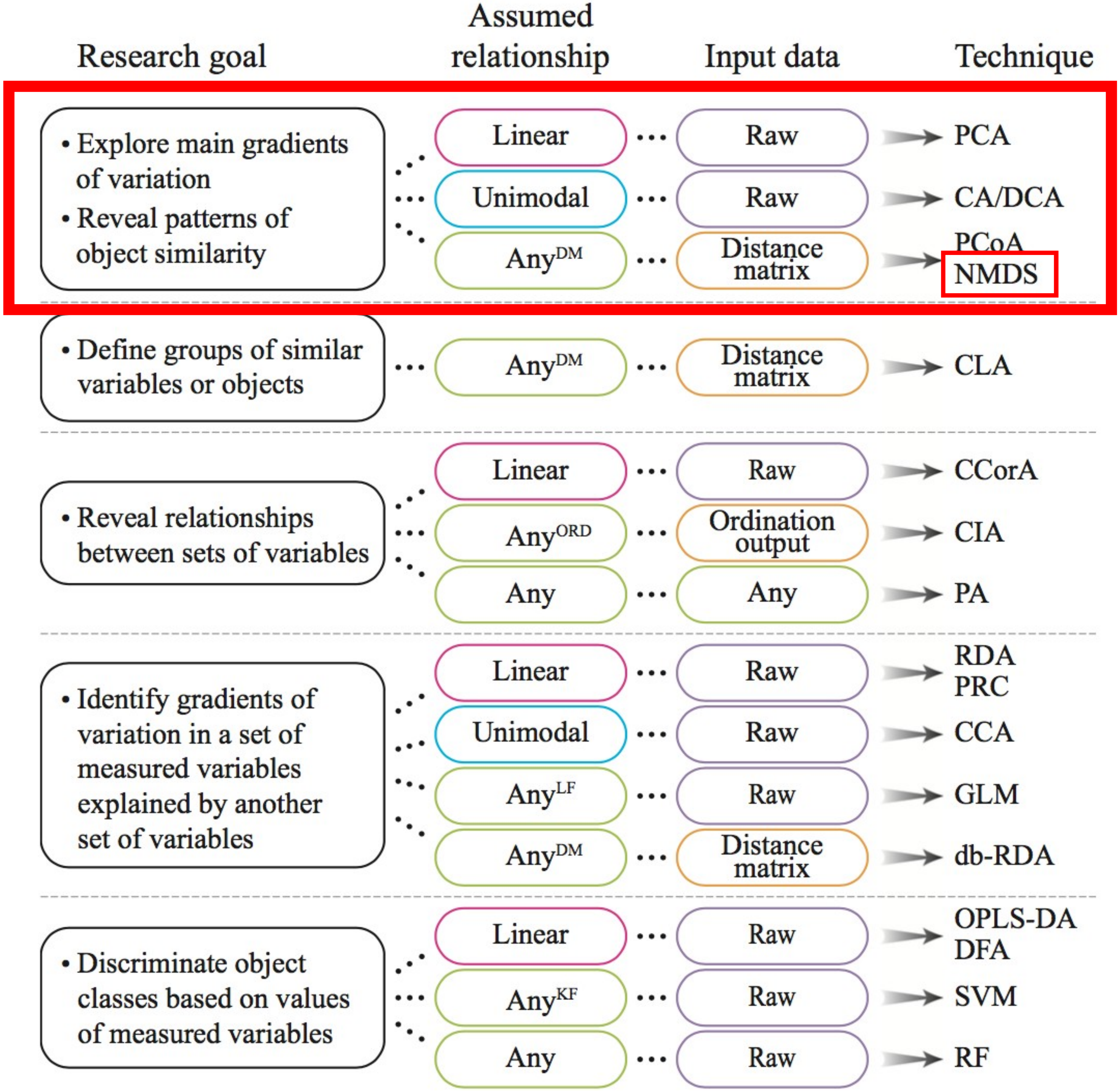


Ordination

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



Unconstrained ordination with NMDS



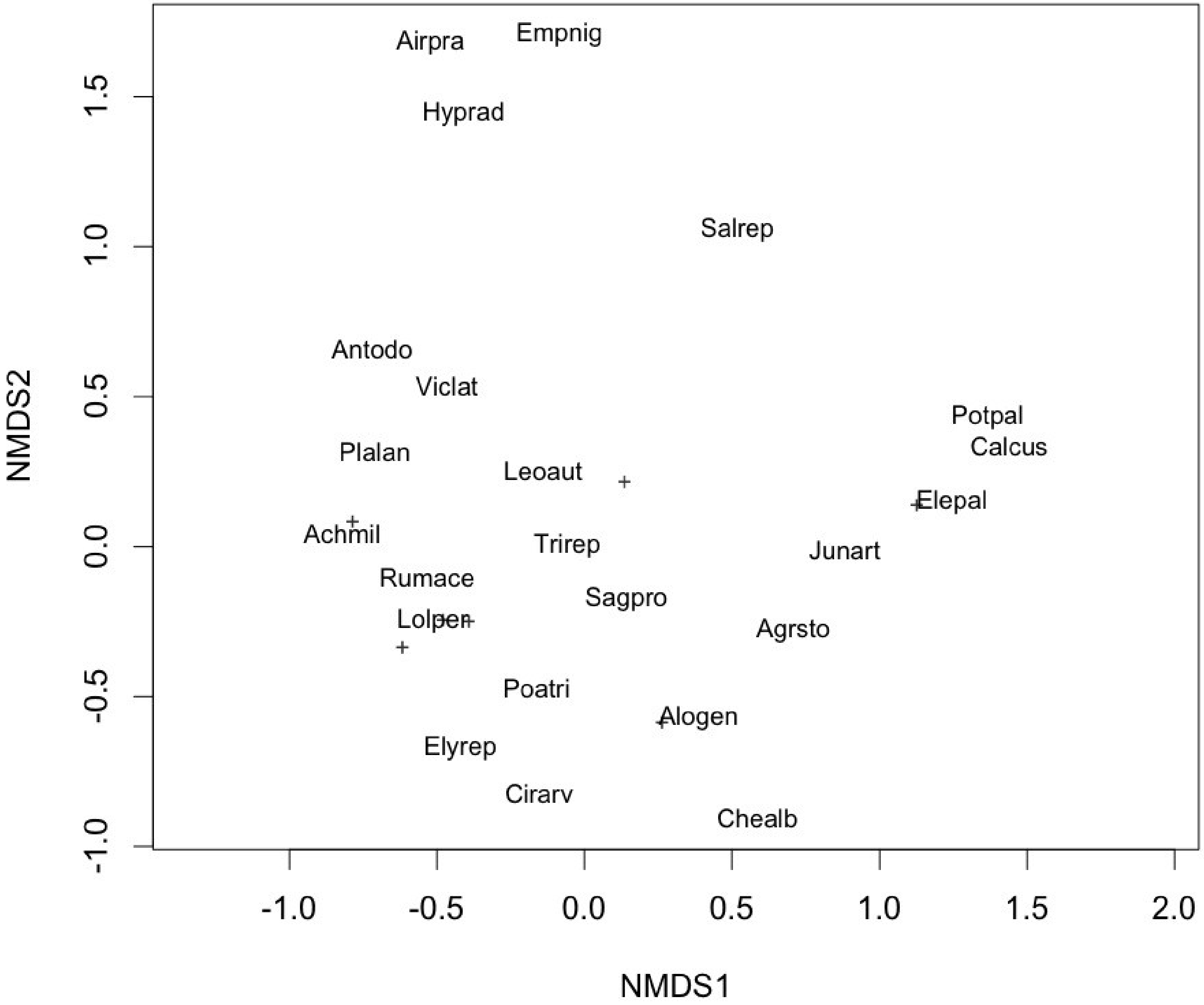
Research goal	Assumed relationship	Input data	Technique
• Explore main gradients of variation • Reveal patterns of object similarity	Linear	Raw	PCA
	Unimodal	Raw	CA/DCA
	Any^{DM}	Distance matrix	PCoA NMDS
• Define groups of similar variables or objects	Any^{DM}	Distance matrix	CLA
• Reveal relationships between sets of variables	Linear	Raw	CCorA
	Any^{ORD}	Ordination output	CIA
	Any	Any	PA
• Identify gradients of variation in a set of measured variables explained by another set of variables	Linear	Raw	RDA PRC
	Unimodal	Raw	CCA
	Any^{LF}	Raw	GLM
	Any^{DM}	Distance matrix	db-RDA
• Discriminate object classes based on values of measured variables	Linear	Raw	OPLS-DA DFA
	Any^{KF}	Raw	SVM
	Any	Raw	RF

Paliy & Shankar 2016 *Mol Ecol*:1032

Non-metric multidimensional scaling

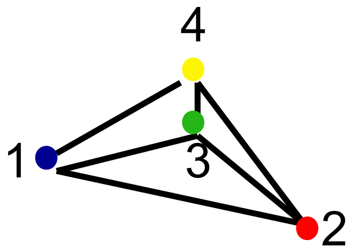
- Unconstrained ordination for different distance metrics, based on ordered distances

- Suitable for ecological data
- Not based on eigenvalues, no partitioning of variance
- Very robust and flexible

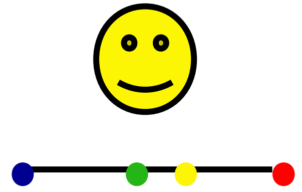


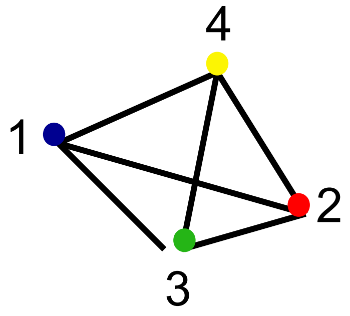
Understanding NMDS

The challenge of visualising distances in a lower dimension:

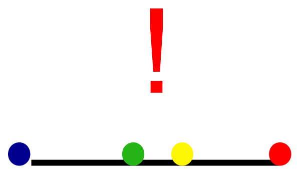


$$\begin{pmatrix} 0 & . & . & . \\ 1.3 & 0 & . & . \\ 0.4 & 0.6 & 0 & . \\ 0.5 & 0.7 & 0.1 & 0 \end{pmatrix}$$





$$\begin{pmatrix} 0 & . & . & . \\ 1.3 & 0 & . & . \\ 0.4 & 0.6 & 0 & . \\ 0.5 & 1.3 & 1.1 & 0 \end{pmatrix}$$



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

Steps of NMDS algorithm

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

From distance to disparity matrix

Distance matrix for data

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

Ordered distances of distance matrix

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

Distance matrix of the initial configuration

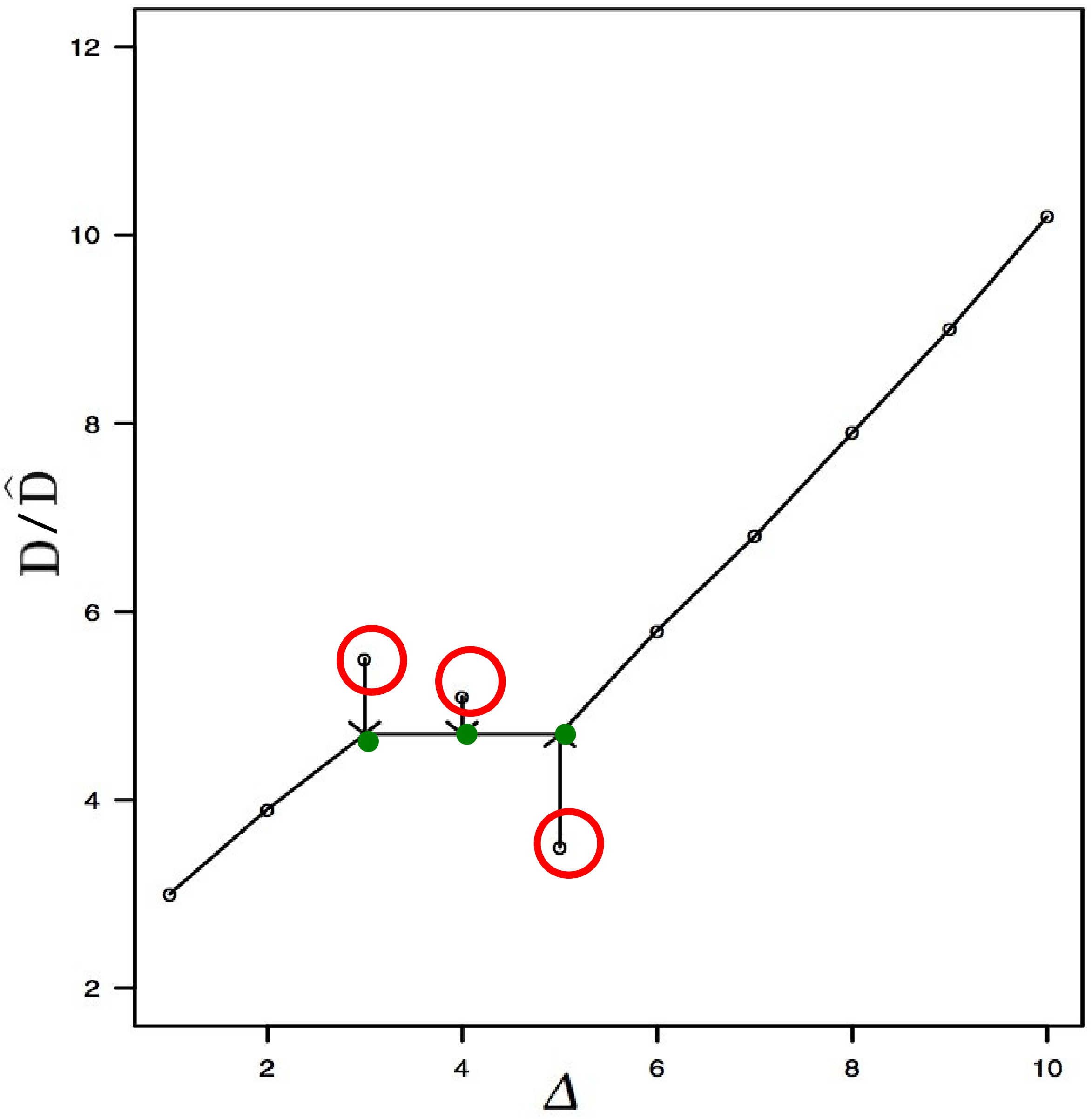
$$\mathbf{D} = \begin{pmatrix} 0 & 9.0 & \boxed{5.1} & 10.2 & 6.8 \\ 9.0 & 0 & \boxed{5.5} & 3.0 & 3.9 \\ 5.1 & 5.5 & 0 & 7.9 & 5.8 \\ 10.2 & 3.0 & 7.9 & 0 & \boxed{3.5} \\ 6.8 & 3.9 & 5.8 & 3.5 & 0 \end{pmatrix}$$

Monotone regression

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

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

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



Handl 2010: 174

From distance to disparity matrix

Distance matrix for data

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

Ordered distances of distance matrix

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

Ordered distances of disparity matrix

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

Disparity matrix

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

Goodness of fit for NMDS

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

Value of *STRESS1*	Goodness of configuration
< *0.05*	excellent
< *0.10*	good
< *0.15*	medium
> 0.15	bad

Implementation of NMDS in R

`monoMDS() {vegan}` Basic function for NMDS

`metaMDS() {vegan}` „Shotgun“ method

`cmdscale() {stats}`

`cmds() {mclust}` (Metric) multidimensional scaling

What does the “shotgun” method do?

`metaMDS() {vegan}`

1. Data transformation (Square-root and Wisconsin double transformation)

2. Calculation of distance matrix based on the selected similarity coefficient (defaults to Bray-Curtis)

3. Adjustment in no shared occurrences of species

4. Several random starts for initial configuration

5. Centring and rotation of ordination (highest dispersion on 1st axis)

6. Scaling (1 unit means halving of community similarity)

7. Calculation of species scores as weighted averages of sites

Limitations of NMDS

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