Multivariate statistics in R

Hannes PETER
Martin BOUTROUX
Zhe Liu

discussion points

- Evaluation?
- Group work?
- Notebooks?
- Questions for last week?

recap – session 1

- Multivariate data
- Use R
- Data exploration
  - Data import, data processing
  - Summary/descriptive statistics
  - Visualizations
    - histograms
    - heatmaps

Paper discussion

Annu. Rev. Ecol. Syst. 1990. 21:129–66

MULTIVARIATE ANALYSIS IN ECOLOGY AND SYSTEMATICS: PANACEA OR PANDORA'S BOX?

Frances C. James

Department of Biological Science, Florida State University, Tallahassee, Florida 32306

Charles E. McCulloch

Biometrics Unit, Cornell University, Ithaca, New York 14853

KEY WORDS: multivariate analysis, data analysis, statistical methods

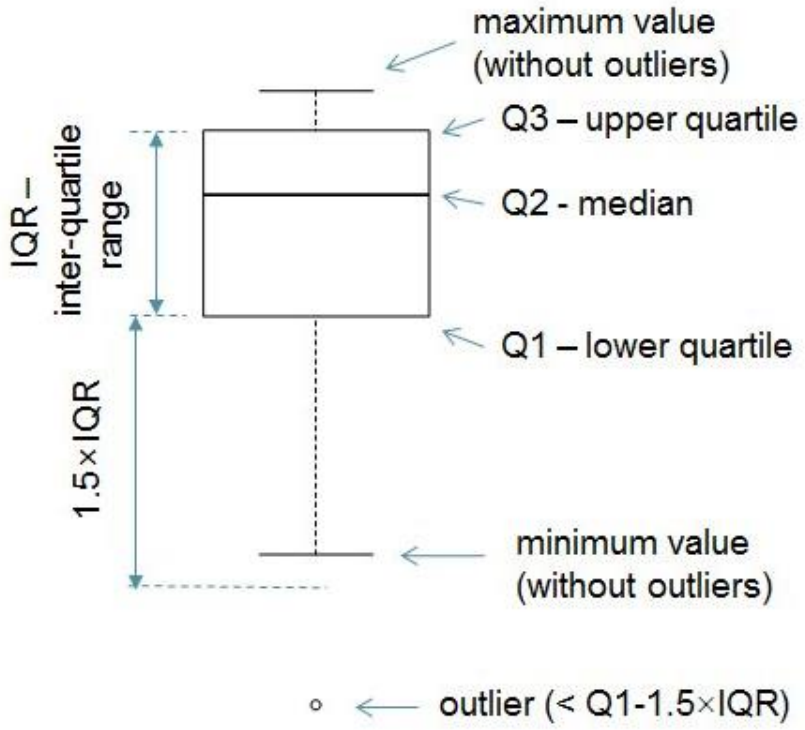
EPFL

Transformation

Association metrics

Data exploration

- Use descriptive statistics (range, median, etc,…) to explore data
- Visualize raw data (histograms, boxplots)
- Check for outliers
- Check for missing values (NA)
  - NA != 0
- Transformation?



Transformation

Mathematical operations applied to the data to change (relative) differences and distributions.

=> make descriptors comparable (e.g. pH, temperature, elevation,…)

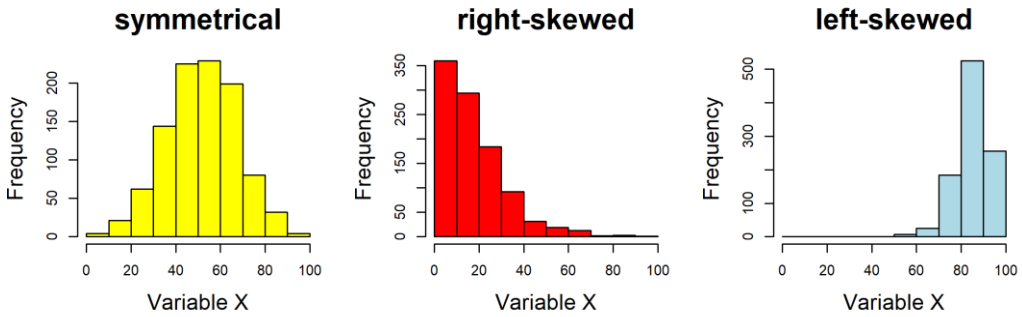
=> statistical tests can be sensitive to distribution (of residuals) (e.g. normal distribution) and variance (homoscedasticity).

=> linear relationships are easier to interpret than non-linear ones

=> emphasize the importance of rare (low-abundance) taxa

Transformation

- semi-quantitative to quantitative descriptors
  - ex.: dominance, ranks into coverage
- quantitative to binary descriptors
  - ex.: species abundance into presence-absence
- quantitative to qualitative or semi-quantitative descriptors
  - ex.: transformation into classes
- improve descriptor distribution
  - linear transformations
    - adding a constant
  - non-linear transformations
    - square root (slight right skew)
    - logarithm (strong right skew)
    - power (left skew)
    - arcsine (proportions)



Example:

Transformations of Braun-Blanquet codes of vegetation cover

Quantitative transformations based on ranks $y = rank^w$

Codes abundance/ dominance	Central values of classes, in % of cover	Ranks	w=0	w=0.25	w=0.5	w=1	w=2	w=4
absent	0.0	0	0.0	0.00	0.00	0	0	0
rare	( )	1	1.0	1.00	1.00	1	1	1
+	0.1	2	1.0	1.19	1.41	2	4	16
1	5.0	3	1.0	1.32	1.73	3	9	81
2m		4	1.0	1.41	2.00	4	16	256
2 2a	17.5	5	1.0	1.50	2.24	5	25	625
2b		6	1.0	1.57	2.45	6	36	1296
3	37.5	7	1.0	1.63	2.65	7	49	2401
4	62.5	8	1.0	1.68	2.83	8	64	4096
5	87.5	9	1.0	1.73	3.00	9	81	6561

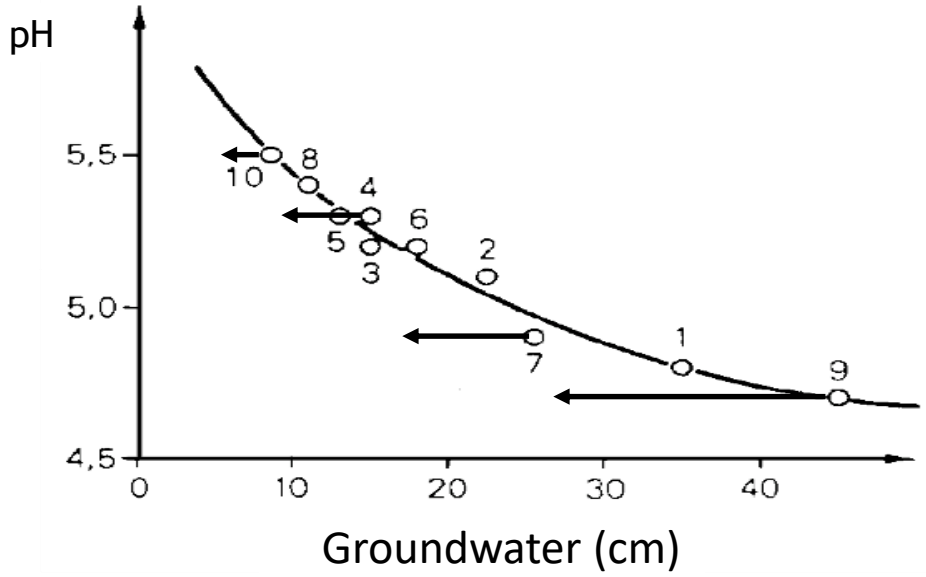
Presence/absence

Approx. % original

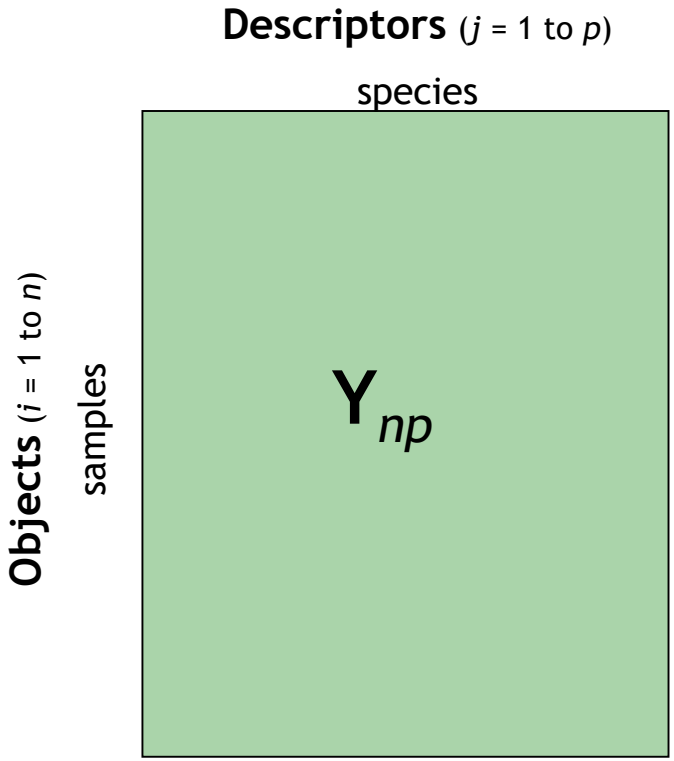
Example:
Data transformation to improve linearity

log-transformation of groundwater depth allows to pass from a curvilinear relation (left) to a linear relation (right) of this variable with pH

Ste	1	2	3	4	5	6	7	8	9	10
pH	4.8	5.1	5.2	5.3	5.3	5.2	4.9	5.4	4.7	5.5
Groundwater	35	22	15	15	13	18	26	11	45	9
Groundwater, log	1.54	1.34	1.18	1.18	1.11	1.26	1.41	1.04	1.65	0.95



Transformation of **descriptors** or **objects**



Transformation of descriptors (species)

- Scaling of values between 0 and 1 or between -1 and 1 (*ranging*)

$$y'_i = \frac{y_i}{y_{\max}} \qquad y'_i = \frac{y_i - y_{\min}}{y_{\max} - y_{\min}}$$

- Centering and reduction (*standardization, z-scores*)
  - Mean= 0
  - Standard deviation = 1

$$z_i = \frac{y_i - \bar{y}}{s_y}$$

- Transformation into *relative values* (proportions per descriptor, species profiles)
  - Total per descriptor = 1
  - Species' profile : differences of abundance between species of the same community are not preserved
  - Good choice when the focus is on species and comparing their ecological niche

$$y'_{ij} = \frac{y_{ij}}{\sum_{i=1}^{n} y_{ij}} = \frac{y_{ij}}{y_{+j}}$$

Transformation of objects (samples)

- Transformation into *relative values* (proportions per object, *site profiles*)
  - Total per object = 1
  - Site profile: differences of abundances of one species across various sites are not preserved
  - Often best choice when focus is on species assemblages in biomonitoring (single sites, no cross-site comparison)

$$y'_{ij} = \frac{y_{ij}}{\sum_{j=1}^{p} y_{ij}} = \frac{y_{ij}}{y_{i+}}$$

- *Hellinger transformation*
  - Recommended in case of many absences in a species matrix

$$y'_{ij} = \sqrt{\frac{y_{ij}}{y_{i+}}}$$

- *Normalization* of object vectors
  - Every value is divided by the norm (length) of the object vector
  - The norm of every object vector is adjusted to 1

$$y'_{ij} = \frac{y_{ij}}{\sqrt{\sum_{j=1}^{p} y_{ij}^2}}$$

Hellinger transformation

Samples	Species			$\sum$
	sp 1	sp 2	sp 3	
sample 1	$y_{11} = 0$	$y_{12} = 1$	$y_{13} = 1$	$y_{1+} = 2$
sample 2	$y_{21} = 1$	$y_{22} = 0$	$y_{23} = 0$	$y_{2+} = 1$
sample 3	$y_{31} = 0$	$y_{32} = 4$	$y_{33} = 8$	$y_{3+} = 12$

$$y'_{ij} = \sqrt{\frac{y_{ij}}{y_{i+}}}$$

$\downarrow \frac{y_{ij}}{y_{i+}}$

Samples	Species		
	sp 1	sp 2	sp 3
sample 1	0	0.5	0.5
sample 2	1	0	0
sample 3	0	0.33	0.67

$\rightarrow \sqrt{\ }$

Samples	Species		
	sp 1	sp 2	sp 3
sample 1	0	0.707	0.707
sample 2	1	0	0
sample 3	0	0.577	0.816

Reduces differences in absolute abundances between samples and reduces the effect of species with high abundances.

- Euclidean distance of Hellinger-transformed species abundance (= Hellinger distance) is well suited for ordination!

Special case transformations

- *Double standardization* (joint standardization of objects and descriptors)
  - *Chi-square (X^2) transformation*

$$y'_{ij} = \frac{y_{ij}}{y_{i+}\sqrt{y_{+j}}}$$

  - *Wisconsin-double transformation*
    - The descriptors are first scaled between 0 and 1, then objects are transformed into site profiles

$$y'_{ij} = \sqrt{y_{++}}\,\frac{y_{ij}}{y_{i+}\sqrt{y_{+j}}}$$

- *Centered log ratio* (clr) and associated transformations (robust clr, additive log ratio (alr)).


Microbiome Datasets Are Compositional: And This Is Not Optional

Gregory B. Gloor[1]*, Jean M. Macklaim[1], Vera Pawlowsky-Glahn[2] and Juan J. Egozcue[3]

[1] Department of Biochemistry, University of Western Ontario, London, ON, Canada, [2] Departments of Computer Science, Applied Mathematics, and Statistics, Universitat de Girona, Girona, Spain, [3] Department of Applied Mathematics, Universitat Politècnica de Catalunya, Barcelona, Spain

It is good practice to check the result of transformation by producing a plot or computing summary statistics

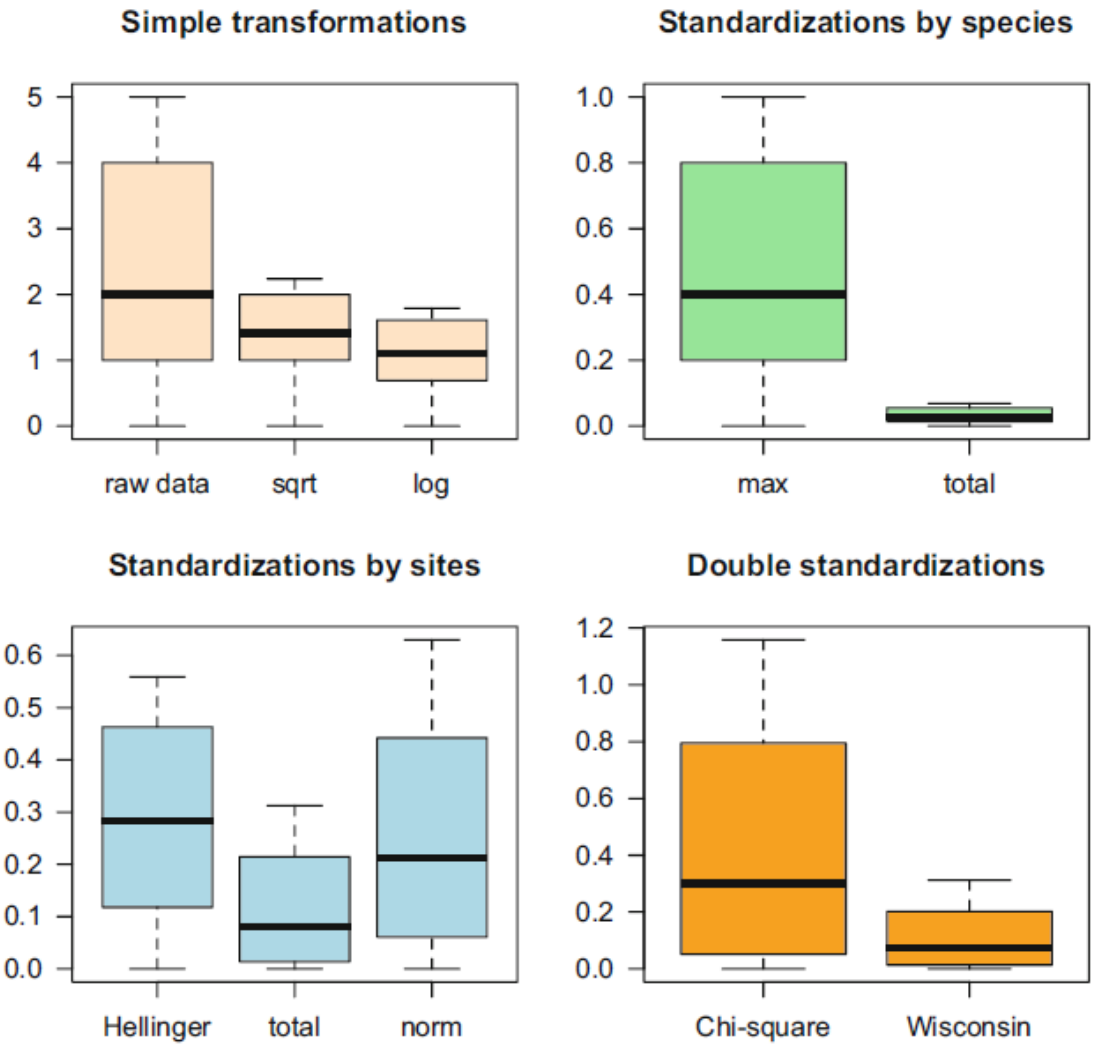
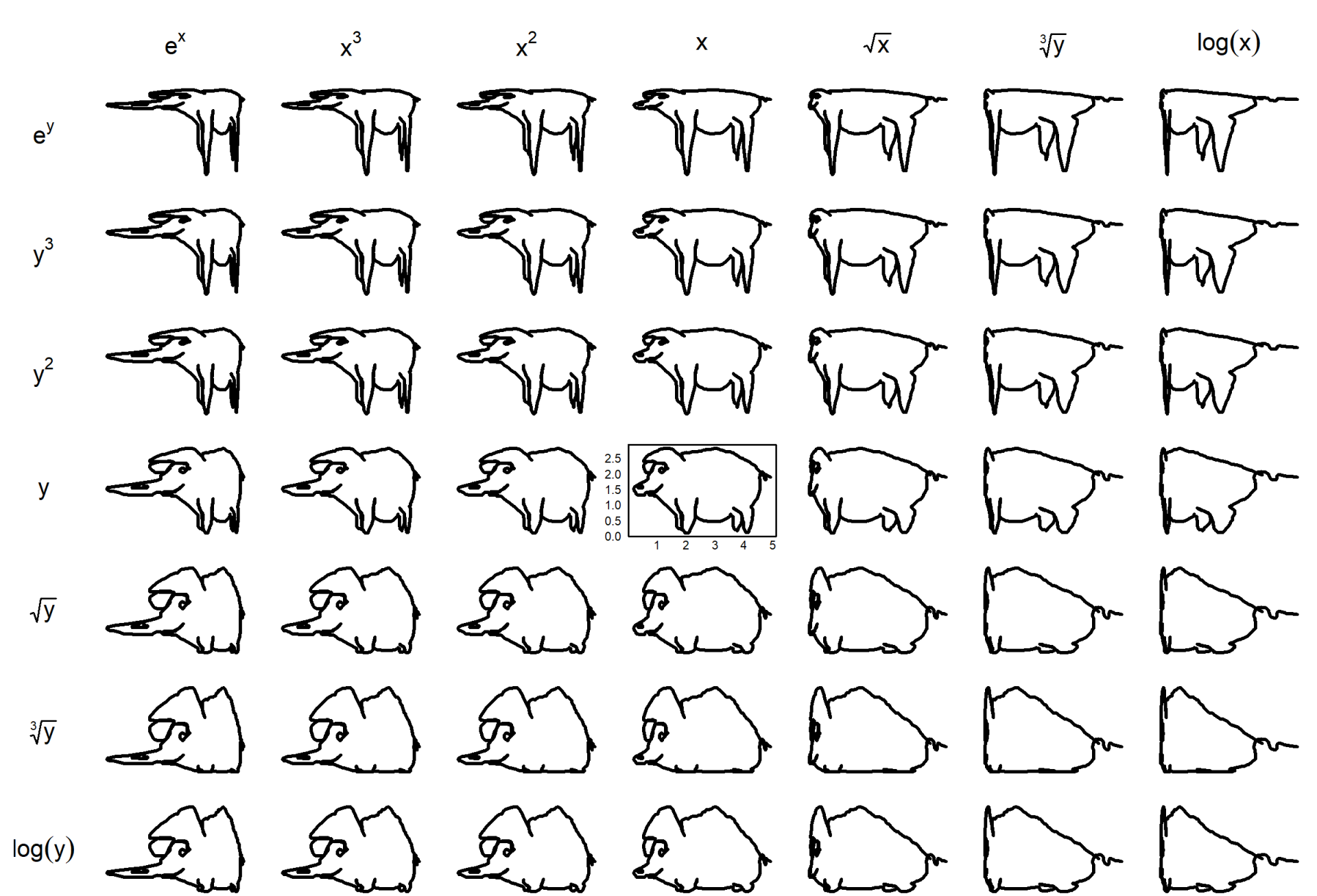


Fig. 2.6 Boxplots of transformed abundances of a common species, *Barbatula barbatula* (stone loach)



https://www.davidzeleny.net/

Association Measures

Similarity (distance) between objects

Interdependence (correlation) of descriptors

Origin and the consequences of the double zero problem

Association Measures

Most methods of multivariate analysis, in particular ordination and clustering techniques, are based on the comparison of all possible pairs of objects or descriptors.

=> Multivariate analyses are done on association matrices, the **choice of an appropriate measure is crucial**.

Choice depends on:

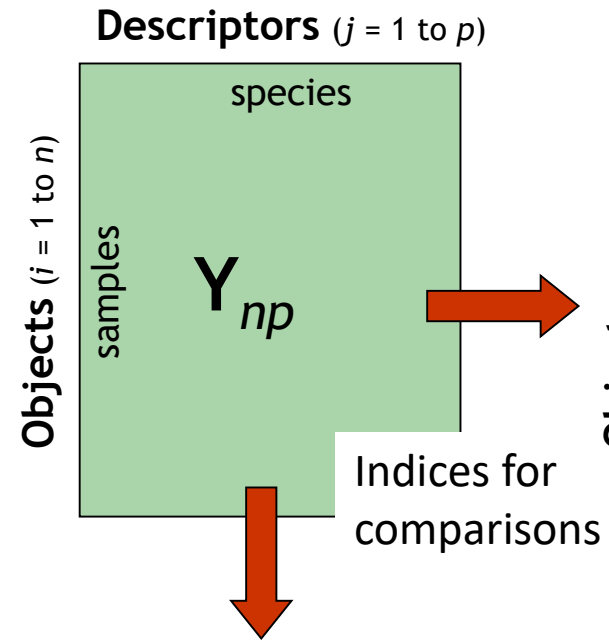
- research question
- type of comparison (objects or descriptors)
- type and mathematical property of variables (species or physicochemical, quantitative, qualitative,…)

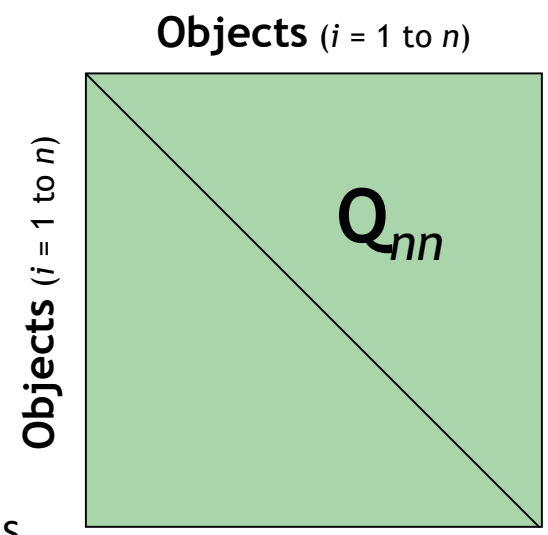
Comparisons take the form of **association measures** (often called coefficients or indices), which are assembled in an **association matrix**.

Q- and R- mode

Association (resemblance) matrix

Rectangular matrix of raw data (Y)





Q mode

comparison between pairs of samples

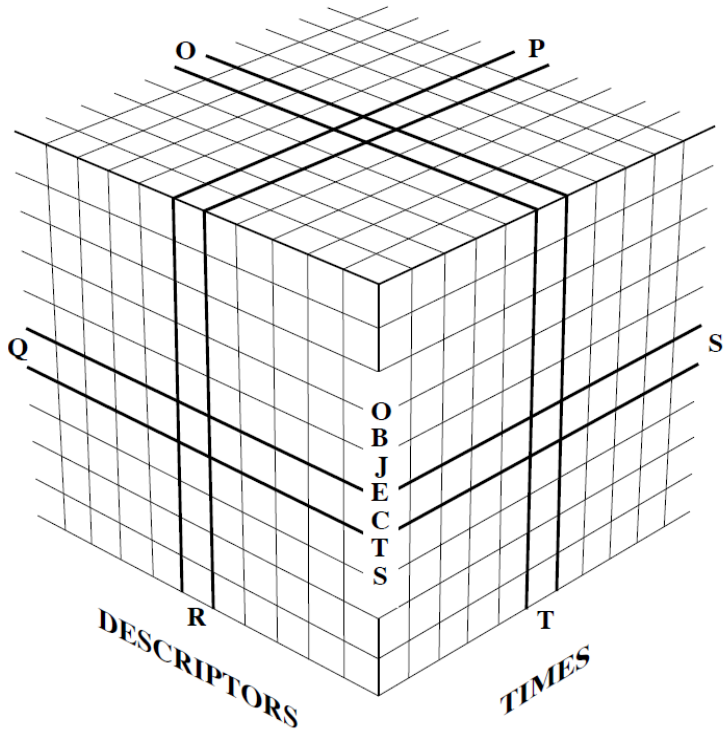
⇒ **similarity/ dissimilarity**

R mode

comparison between pairs of species

⇒ **dependence**

(correlation)



The three-dimensional data box (objects × descriptors × times). Adapted from Cattell (1966).

O: among time instances, based on all observed descriptors (a single object);

P: among descriptors, based on all observed times (a single object);

Q: among objects, based on all observed descriptors (a single instance);

R: among descriptors, based on all observed objects (a single instance);

S: among objects, based on all observed times (a single descriptor);

T: among time instances, based on all observed objects (a single descriptor).

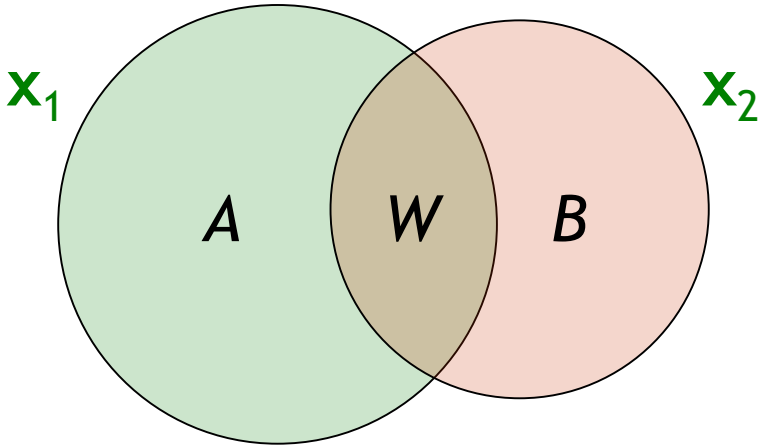
Association measures (Q mode)

- Measure of similarity $S(\mathbf{x}_1,\mathbf{x}_2)$
  - between 0 and 1
  - 1 for two identical objects
- Measure of dissimilarity $D(\mathbf{x}_1,\mathbf{x}_2)$
  - between 0 and 1
  - 0 for two identical objects
- Measure of distance $D(\mathbf{x}_1,\mathbf{x}_2)$
  - No supremum (or > 1)
  - 0 for two identical objects

- Measures of similarity can be converted to dissimilarity and reciprocally
  $D = 1 - S$ $\qquad\qquad$ $S = 1 - D$

Similarity (Q mode)

Similarity indices for **binary descriptors** (presence-absence)



Venn notation

- W = number of species in common
- Different weights of shared/non-shared species

Jaccard $$S_7^*(\mathbf{x}_1, \mathbf{x}_2) = \frac{W}{A + B - W}$$

Sørensen $$S_8(\mathbf{x}_1, \mathbf{x}_2) = \frac{2W}{A + B}$$

Ochiai $$S_{14}(\mathbf{x}_1, \mathbf{x}_2) = \frac{W}{\sqrt{AB}}$$

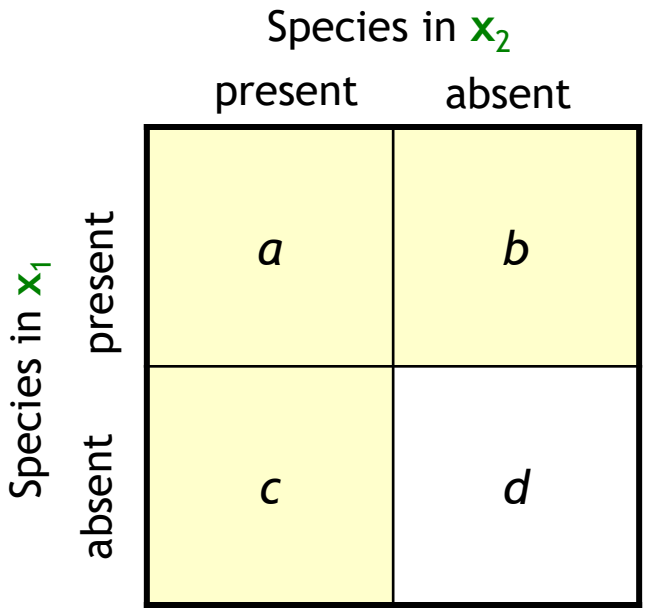
Kulczyński $$S_{18}(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{2}\left(\frac{W}{A} + \frac{W}{B}\right)$$

** numbering of similarity and dissimilarity measures (S_7, S_8 etc.) according to Legendre and Legendre (2012).*

Similarity (Q mode)

Notation by contingency table

d = « double zeros »



	Similarity	Dissimilarity (1 – S)
Jaccard	$S_7(\mathbf{x}_1,\mathbf{x}_2)=\frac{a}{a+b+c}$	$D(\mathbf{x}_1,\mathbf{x}_2)=\frac{b+c}{a+b+c}$
Sørensen	$S_8(\mathbf{x}_1,\mathbf{x}_2)=\frac{2a}{2a+b+c}$	$D_{13}(\mathbf{x}_1,\mathbf{x}_2)=\frac{b+c}{2a+b+c}$

Example

	Survey x1	Survey x2	
Species 1	1	1	*a*
Species 2	0	0	*d*
Species 3	0	0	*d*
Species 4	1	1	*a*
Species 5	1	0	*b*
Species 6	1	0	*b*
Species 7	0	1	*c*
Species 8	1	1	*a*
Total	**5**	**4**	

Venn notation →

A	5 species in survey 1
B	4 species in survey 2
W	3 species in common

contingency table →

a	3	species found in both surveys
b	2	species found only in survey 1
c	1	species found only in survey 2
d	2	species missing in both surveys

Jaccard Similarity

$$S_7(\mathbf{x}_1, \mathbf{x}_2) = \frac{a}{a+b+c} \qquad \frac{3}{3+2+1}$$

Jaccard Dissimilarity

$$D(\mathbf{x}_1, \mathbf{x}_2) = \frac{b+c}{a+b+c} \qquad \frac{2+1}{3+2+1}$$

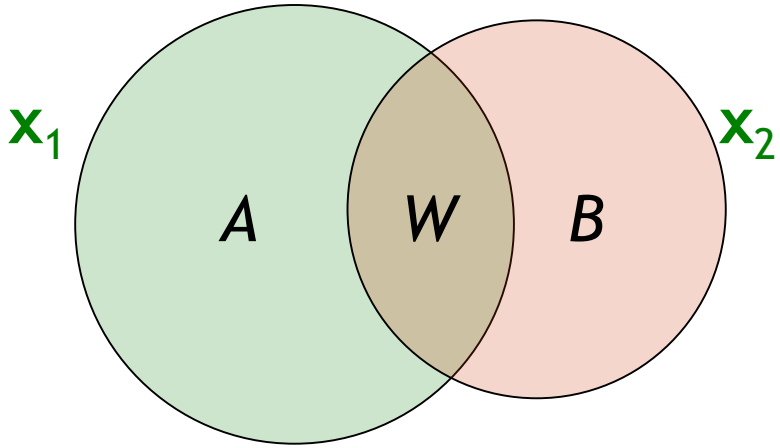
Similarity Indices ↓

Jaccard	**0.500**
Sørensen	**0.667**
Ochiai	**0.671**
Kulczyński	**0.675**
Sokal-Michener	**0.625**

Similarity (Q mode)

Quantitative or semi-quantitative descriptors (abundance, dominance, frequency, etc.)

- W is assessed differently: sum of the smallest values of abundance of shared species



	Similarity	Dissimilarity (1 – S)
Jaccard	$S(\mathbf{x}_1,\mathbf{x}_2)=\frac{W}{A+B-W}$ Van der Maarel	$D(\mathbf{x}_1,\mathbf{x}_2)=\frac{A+B-2W}{A+B-W}$ Ruzicka
Sørensen	$S_{17}(\mathbf{x}_1,\mathbf{x}_2)=\frac{2W}{A+B}$ Steinhaus	$D_{14}(\mathbf{x}_1,\mathbf{x}_2)=\frac{A+B-2W}{A+B}$ Bray-Curtis*

*Odum's percentage difference

Example

Species abundance:

	Survey x1	Survey x2	shared (min)
Species 1	10	5	5
Species 2	0	0	0
Species 3	0	0	0
Species 4	60	10	10
Species 5	5	0	0
Species 6	5	0	0
Species 7	0	10	0
Species 8	10	5	5
Sum	**90**	**30**	**20**

$$\frac{90 + 30 - 2 \times 20}{90 + 30} = 0.66$$

$$D_{14}(\mathbf{x}_1, \mathbf{x}_2) = \frac{A + B - 2W}{A + B}$$

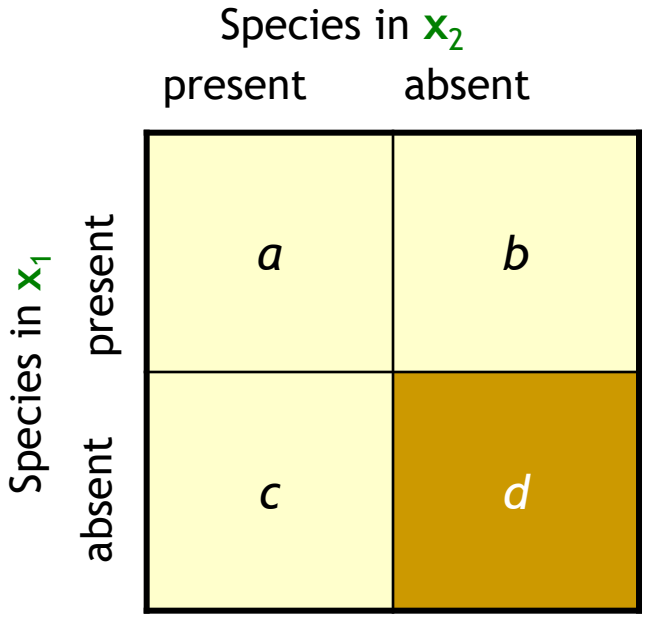
Bray-Curtis dissimilarity

A	90	Sum of species of survey 1
B	30	Sum of species of survey 2
W	20	Sum of species in common (minima)

double zero problem

The absence of species should not contribute to similarity between samples.

- Symmetrical indices allow for double zeros and therefore should be avoided when comparing lists of species, but are OK if the zero has an unambiguous meaning (ex. 0 mg/L O_2)
  - Ex. index of simple concordance (Sokal-Michener)
- Asymmetrical indices (Jaccard, Sørensen, …) should be used for species presence/absence or abundance

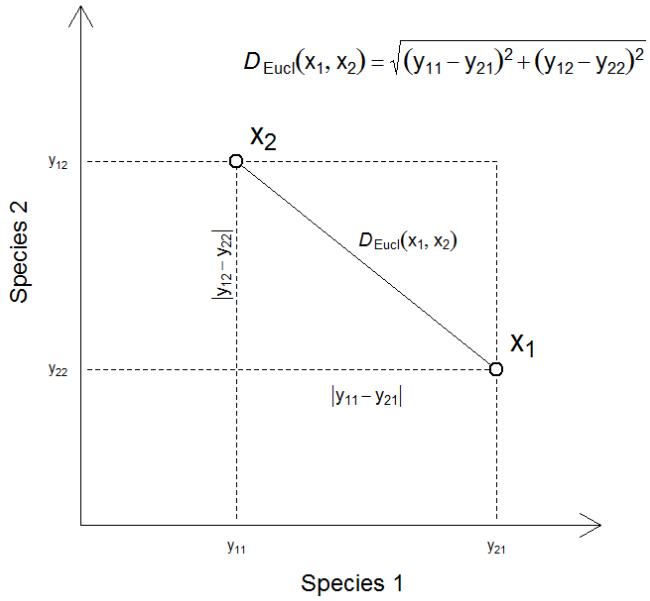


Sokal-Michener

$$S_1(\mathbf{x}_1,\mathbf{x}_2) = \frac{a+d}{p} \qquad p=a+b+c+d$$

$$D(\mathbf{x}_1,\mathbf{x}_2) = \sqrt{1 - S_1(\mathbf{x}_1,\mathbf{x}_2)} = \sqrt{\frac{b+c}{p}}$$

double zero problem



	Species 1	Species 2	Species 3
Sample 1	0	1	1
Sample 2	1	0	0
Sample 3	0	4	8

$$D_{Eucl}\left(Sample\,1, Sample\,2\right) = \sqrt{\left(0-1\right)^2+\left(1-0\right)^2+\left(1-0\right)^2} = 1.732$$

$$D_{Eucl}\left(Sample\,1, Sample\,3\right) = \sqrt{\left(0-0\right)^2+\left(1-4\right)^2+\left(1-8\right)^2} = 7.615$$

Coefficients which skip double zeros are called *asymmetrical* because they treat double absences in a different way than double presences.

Distance between objects (Q mode)

Euclidean distance

- Used in Principal Component Analysis (PCA)
- No upper limit
- Increases with the number of descriptors

Chord distance

- Euclidean distance of normalized objects
- Bound between 0 and $2^{0.5}$
- Does not increase with the number of descriptors

Manhattan Distance

- No upper limit
- Depends on the scale of variables
- Calculated preferably after standardization of descriptors

Example

Square root of the sum of squares of the values

	y_1	y_2	y_3	y_4	y_5	Norm
x_1	45	10	15	0	10	49.5
x_2	25	8	10	0	3	28.2
x_3	7	15	20	14	12	31.8

45/49.5

Object vectors normed

	y_1	y_2	y_3	y_4	y_5
x_1	0.909	0.202	0.303	0.000	0.202
x_2	0.885	0.283	0.354	0.000	0.106
x_3	0.220	0.471	0.628	0.440	0.377

Euclidean distance

$$D_1(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{(45-25)^2 + \ldots + (0-0)^2 + (10-3)^2} = 21.9$$

Chord distance

$$D_3(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{(0.909-0.885)^2 + \ldots + (0-0)^2 + (0.202-0.106)^2} = 0.138$$

Manhattan distance

$$D_7(\mathbf{x}_1, \mathbf{x}_2) = |45-25| + \ldots + |0-0| + |10-3| = 34$$

Distance between objects (Q mode)

- χ^2 Distance
  - Euclidean distance of χ^2 transformed data
    (divide by row sums and square root of column sums, and adjusted for square root of matrix total)
  - Used in correspondence analysis (CA)
  - No upper limit

$$D_{16}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{p} \frac{y_{++}}{y_{+j}} \left(\frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}} \right)^2}$$

- Hellinger Distance
  - Variant of the χ^2 metric
  - Euclidean distance on Hellinger transformed data (square root of value divided by total)
  - Better than the preceding ones for linear ordination (e.g. PCoA)
  - No upper limit

$$D_{17}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{p} \left(\sqrt{\frac{y_{1j}}{y_{1+}}} - \sqrt{\frac{y_{2j}}{y_{2+}}} \right)^2}$$

Special-case distance metrics

- Gower's distance
  - combinations of logical, numerical, categorical or text data
- Raup-Crick
  - probabilistic index based on presence/absence data. It is defined as 1 - prob(j), or based on the probability of observing at least j shared species in compared communities.

Measures of dependence between descriptors (R mode)

main purpose of R-mode analysis is to investigate relationships among descriptors

R mode dependence matrices may also be used, in some cases, as the computational basis for the ordination of *objects*, e.g. in principal component or linear discriminant analyses

Measures of dependence between descriptors (R mode) (parametric)

- Covariance
  - Measuring the joint dispersion of two variables around their means
  - No lower bound or supremum

$$\mathrm{cov}(\mathbf{y}_1, \mathbf{y}_2) = \frac{\sum_{i=1}^{n} (y_{i1} - \bar{y}_1)(y_{i2} - \bar{y}_2)}{n-1}$$

- Pearson Correlation *r*
  - between -1 and 1

$$r(\mathbf{y}_1, \mathbf{y}_2) = \frac{\sum_{i=1}^{n} (y_{i1} - \bar{y}_1)(y_{i2} - \bar{y}_2)}{\sqrt{\sum_{i=1}^{n} (y_{i1} - \bar{y}_1)^2 \sum_{i=1}^{n} (y_{i2} - \bar{y}_2)^2}}$$

Dependence measures are non-metric because they can show negative values.

They are testable (p-value, check assumptions).

Parametric dependence only for linear relationships!

Measures of dependence between descriptors (R mode) (non-parametric)

Spearman rank correlation (rho)

- between -1 and 1

$$\rho(\mathbf{y}_1, \mathbf{y}_2) = 1 - \frac{6\sum_{i=1}^{n}(y_{i1} - y_{i2})^2}{n^3 - n}$$

Kendall rank correlation (tau)

- between -1 and 1
- S= sum of concordances (-1 or +1) between ranks of paired descriptors

$$\tau(\mathbf{y}_1, \mathbf{y}_2) = \frac{2S}{n(n-1)}$$

n = total number of objects