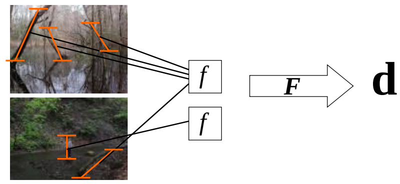
Advanced Community Ecology

Spring 2019, C. Blackwood notes

- Multivariate Methods
  - Step 1. Distance metrics

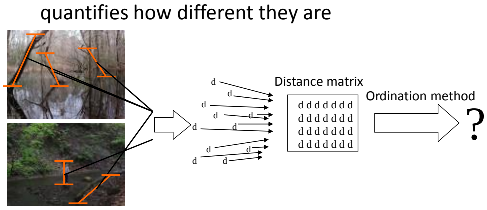
Steps in a multivariate analysis

- 1. **Choose a distance coefficient**
  - Basis for pairwise comparisons
  - Each pair of profiles is given a number which quantifies how different they are



Steps in a multivariate analysis

- 1. **Choose a distance coefficient**
  - Basis for pairwise comparisons
  - Each pair of profiles is given a number which quantifies how different they are

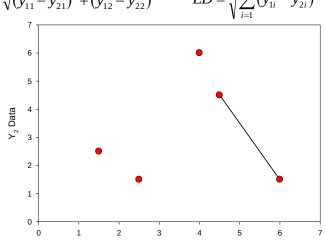


- Results of a multivariate method reflect the distance matrix
  - Not necessarily the original dataset, or what you want to know about the original dataset

Euclidean Distance

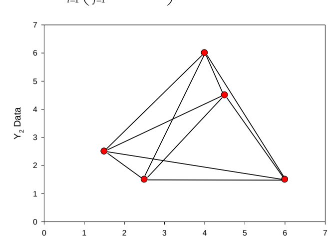
- Standard mathematics and geometry are Euclidean

$$ED = \sqrt{(y_{11} - y_{21})^2 + (y_{12} - y_{22})^2} \qquad ED = \sqrt{\sum_{i=1}^{p} (y_{1i} - y_{2i})^2}$$



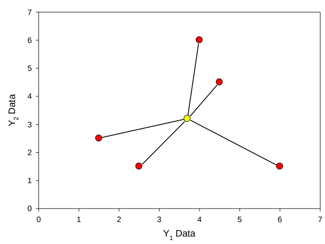
Mean ED^2 in a group of *n* points

$$MeanED^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} (y_{ij} - y_{ij})^2 \right) \qquad Var_j = \frac{1}{n-1} \sum_{i=1}^{n} (y_{ij} - \bar{y}_j)^2$$



Standard statistics is Euclidean

$$MeanVar = \frac{1}{n-1}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\left(y_{ij}-\bar{y}_j\right)^2\right) = \frac{MeanED^2}{2}$$



Distance Coefficient Assumptions

1. Geometric properties

- Semimetric coefficients
  - 1. Distances are >= 0
  - 2. Distance does not depend on direction
- Metric coefficients
  - 1. Satisfy semimetric properties
  - 2. Triangle inequality: The *sum* of the distances between points (A,C) and points (B,C) is >= the distance between points (A,B)
    i.e. The distance from A to B is the shortest path from A to B
  - Euclidean distance is metric

Distance Coefficient Assumptions

2. Weighting rare vs. common species

- Arguments in favor of weighting rare OTUs heavily
  - OTUs occurring less frequently may be sensitive indicators of system change
- Arguments in favor of downweighting rare OTUs
  - OTUs with greater abundance may be more ecologically important
  - Absence from profile does not imply absence from system
    - Presence of OTUs with small populations may depend on stochastic sampling error
  - OTUs with low abundance may be transients

Distance Coefficient Assumptions

2. Weighting rare vs. common species

- 2 aspects of rarity
  - 1. Frequency of occurrence (proportion of sites)
  - 2. Abundance where detected (population where found)
- Euclidean distance gives very little weight to T-RFs with lower abundance

Distance Coefficient Assumptions

3. Dual absence

- Do we consider the absence of an OTU from two plots a sign of their similarity?
- Asymmetric coefficients do not
  - OTUs could be absent from different samples for different reasons
- Boundedness
- Euclidean distance is symmetric and unbounded

Example: Euclidean Distance

	76 bp	122 bp	157 bp	280 bp	387 bp
Sample A	62	0	1005	0	0
Sample B	0	60	0	777	759
Sample C	0	92	0	2010	2195
Sample D	57	1732	0	0	0

$$ED = \sqrt{\sum_{i=1}^{p}\left(y_{1i}-y_{2i}\right)^2}$$

	Sample A	Sample B	Sample C	Sample D
Sample A	0			
Sample B	1482	0		
Sample C	3143	1893	0	
Sample D	2002	1995	3399	0

Some common coefficients

	Weight of rare species	Binary/ Quantitative	Dual Absence	Metric	Derived from Euclidean	
Euclidean (Raw or Proportional)	Small	Quantitative	Symmetric Unbounded	Yes	Yes	$ERD = \sqrt{\sum_{i=1}^{p}\left(\frac{y_{1i}}{y_{1+}} - \frac{y_{2i}}{y_{2+}}\right)^2}$
Chi-square	Large	Quantitative	Symmetric Unbounded	Yes	Yes	$\chi^2 D = \sqrt{\sum_{i=1}^{p}\frac{1}{y_{+i}}\left(\frac{y_{1i}}{y_{1+}} - \frac{y_{2i}}{y_{2+}}\right)^2}$
Hellinger	Medium	Quantitative	Asymmetric Bounded	Yes	Yes	$HD = \sqrt{\sum_{i=1}^{p}\left(\sqrt{\frac{y_{1i}}{y_{1+}}} - \sqrt{\frac{y_{2i}}{y_{2+}}}\right)^2}$
Jaccard	Large	Binary	Asymmetric Bounded	No	No	$JD = 1 - S_J = 1 - \frac{a}{a+b+c}$
Bray-Curtis	Small	Quantitative	Asymmetric Bounded	No	No	$BCD = 1 - \frac{2\sum_{i=1}^{p} Min(y_{1i}, y_{2i})}{y_{1+} + y_{2+}}$

Other coefficients may be appropriate with other types of data (e.g. multistate unordered data, mixed-type or mixed-scale data)

Example: Euclidean Distance on Relative Proportions

	76 bp	122 bp	157 bp	280 bp	387 bp
Sample A	0.06	0	0.94	0	0
Sample B	0	0.04	0	0.49	0.48
Sample C	0	0.02	0	0.47	0.51
Sample D	0.03	0.97	0	0	0

$$ERD = \sqrt{\sum_{i=1}^{p}\left(\frac{y_{1i}}{y_{1+}} - \frac{y_{2i}}{y_{2+}}\right)^2}$$

	Sample A	Sample B	Sample C	Sample D
Sample A	0			
Sample B	1.16	0		
Sample C	1.17	0.04	0	
Sample D	1.35	1.15	1.17	0

Example: Chi-Square Metric

	76 bp	122 bp	157 bp	280 bp	387 bp
Sample A	0.19	0	0.97	0	0
Sample B	0	0.04	0	0.5	0.48
Sample C	0	0.02	0	0.48	0.51
Sample D	0.11	0.96	0	0	0

$$\chi^2 D = \sqrt{\sum_{i=1}^{p}\frac{1}{y_{+i}}\left(\frac{y_{1i}}{y_{1+}} - \frac{y_{2i}}{y_{2+}}\right)^2}$$

	Sample A	Sample B	Sample C	Sample D
Sample A	0			
Sample B	1.21	0		
Sample C	1.21	0.04	0	
Sample D	1.36	1.15	1.17	0

Example: Hellinger Distance

	76 bp	122 bp	157 bp	280 bp	387 bp
Sample A	0.24	0	0.97	0	0
Sample B	0	0.19	0	0.7	0.69
Sample C	0	0.15	0	0.68	0.71
Sample D	0.18	0.98	0	0	0

$$HD = \sqrt{\sum_{i=1}^{p}\left(\sqrt{\frac{y_{1i}}{y_{1+}}} - \sqrt{\frac{y_{2i}}{y_{2+}}}\right)^2}$$

	Sample A	Sample B	Sample C	Sample D
Sample A	0			
Sample B	1.41	0		
Sample C	1.41	0.06	0	
Sample D	1.38	1.27	1.31	0

Example: Bray-Curtis Distance

	76 bp	122 bp	157 bp	280 bp	387 bp
Sample A	0.06	0	0.94	0	0
Sample B	0	0.04	0	0.49	0.48
Sample C	0	0.02	0	0.47	0.51
Sample D	0.03	0.97	0	0	0

$$BCD = 1 - \frac{2\sum_{i=1}^{p} Min(y_{1i}, y_{2i})}{y_{1+} + y_{2+}}$$

	Sample A	Sample B	Sample C	Sample D
Sample A	0			
Sample B	1	0		
Sample C	1	0.02	0	
Sample D	0.97	0.96	0.98	0

Example: Jaccard Distance

	76 bp	122 bp	157 bp	280 bp	387 bp
Sample A	1	0	1	0	0
Sample B	0	1	0	1	1
Sample C	0	1	0	1	1
Sample D	1	1	0	0	0

$$JD = 1 - S_J = 1 - \frac{a}{a+b+c}$$

	Sample A	Sample B	Sample C	Sample D
Sample A	0			
Sample B	1	0		
Sample C	1	0	0	
Sample D	0.67	0.75	0.75	0

Distance Coefficient Assumptions

4. Quantitative or binary

- Binary coefficients - Account for presence/ absence of OTUs only
  - Give equal weight to OTUs of different abundance
  - Low-abundance species are weighted just like high-abundance species
  - Low-frequency species usually still down-weighted
- Quantitative coefficients
  - Incorporate information on changes in relative abundance of OTUs
  - Weighting of rare species depends on transformations

- Binary similarity coefficients

		Object $\mathbf{x}_2$: 1	Object $\mathbf{x}_2$: 0	
Object $\mathbf{x}_1$	1	a	b	$a+b$
	0	c	d	$c+d$
		$a+c$	$b+d$	$p=a+b+c+d$

Simple Matching

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

Jaccard

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

Class of monotonically related coefficients

Jaccard

Sorenson

$$S_{10}(\mathbf{x}_1, \mathbf{x}_2) = \frac{a}{a+2b+2c} \Rightarrow S_7(\mathbf{x}_1, \mathbf{x}_2) = \frac{a}{a+b+c} \Rightarrow S_8(\mathbf{x}_1, \mathbf{x}_2) = \frac{2a}{2a+b+c}$$

$$S_9(\mathbf{x}_1, \mathbf{x}_2) = \frac{3a}{3a+b+c}$$

- Similarity coefficients for *mixtures of variable types*

Gower similarity

$$S_{15}(\mathbf{x}_1, \mathbf{x}_2) = \frac{\sum_{j=1}^{p} w_{12j} s_{12j}}{\sum_{j=1}^{p} w_{12j}}$$

- Quantitative variables
  $s_{12j} = 1 - [|y_{1j} - y_{2j}| / R_j]$
- Qualitative variables
  s_{12j} = 1 or 0
- Semi-quantitative (ordered) variables
  treat like quantitative, directly or on ranks
- w is a weight that can account for joint absences

Table 7.3 Some properties of the distance coefficients described in Section 7.4.

Distance coefficient	D metric, etc.	D Euclidean	$\sqrt{D}$ metric	$\sqrt{D}$ Euclidean
D_1 (Euclidean distance; eq. 7.32)	metric	Yes	Yes	Yes
D_2 (average distance; eq. 7.34)	metric	Yes	Yes	Yes
D_3 (chord distance; eqs. 7.35, 7.36)	metric	Yes	Yes	Yes
D_4 (geodesic metric; eq. 7.37)	metric	No	Yes	Yes
D_5 (Mahalanobis generalized distance; eq. 7.38)	metric	Yes	Yes	Yes
D_6 (Minkowski metric; eq. 7.43)	metric	*	–	–
D_7 (Manhattan metric; eq. 7.44)	metric	No	Yes	Yes
D_8 (mean character difference; eq. 7.45)	metric	No	Yes	Yes
D_9 (index of association; eqs. 7.47, 7.48)	metric	No	Yes	Yes
D_{10} (Canberra metric; eq. 7.49)	metric	No	Yes	Yes
D_{11} (coefficient of divergence; eq. 7.51)	metric	Yes	Yes	Yes
D_{12} (coefficient of racial likeness; eq. 7.52)	nonmetric	No	No	No
D_{13} (nonmetric coefficient; eq. 7.57)	semimetric	No	Yes	Yes
D_{14} (percentage difference; eq. 7.58)	semimetric	No	Yes	Yes
D_{15} (χ^2 metric; eq. 7.54)	metric	Yes	Yes	Yes
D_{16} (χ^2 distance; eq. 7.55)	metric	Yes	Yes	Yes
D_{17} (Hellinger distance; eq. 7.56	metric	Yes	Yes	Yes
D_{18} (distance between species profiles; eq. 7.53)	metric	Yes	Yes	Yes
D_{19} (modified mean character difference; eq. 7.46)	semimetric	No	No	No

* The result depends on the exponent r.

- Can be extended to quantitative data
- Can be extrapolated from limited sampling

Ecology Letters, (2005) **8**: 148–159 doi: 10.1111/j.1461-0248.2004.00707.x

LETTER

A new statistical approach for assessing similarity of species composition with incidence and abundance data

Anne Chao,[1] Robin L. Chazdon,[2*] Robert K. Colwell[2] and Tsung-Jen Shen[1]

Abstract
The classic Jaccard and Sorensen indices of compositional similarity (and other indices that depend upon the same variables) are notoriously sensitive to sample size, especially for assemblages with numerous rare species. Further, because these indices are based