EEB 698, Fall 2019: Some dissimilarity measures

Goal: one number quantifying dissimilarity in species composition between samples i and j

Data: y_{ik} = abundance (perhaps 0,1) of species k in sample i, so rows = samples
Where needed, Y_i is total abundance in sample i, and
$p_{ik} = y_{ik}/Y_i$, the proportional abundance for species k in sample i

Properties of dissimilarity measures:
1) $d_{ij} \geq 0$. No negative dissimilarities
2) $d_{ij} = 0$ only when two samples are the same, as viewed by the dissimilarity
For some: same abundance, i.e. $y_{ik} = y_{jk}$ for all species (k).
For others: same proportion, i.e. $p_{ik} = p_{jk}$ for all species (k).
3) $d_{ij} = d_{ji}$. Matrix is symmetric. Same dissim. when go from i to j or j to i.

Additional property that defines a distance (not just a dissimilarity):
4) $d_{ik} \leq d_{ij} + d_{jk}$ for 3 samples, i, j and k. Triangle inequality

Measures that satisfy 1-4 are called metric; those that satisfy 1-3 are called semi-metric.

Why this can matter: if metric (4 satisfied), can put the 3 locations on a piece of paper so that distance between them is the dissimilarity. Impossible (triangle doesn't close) when semi-metric.

What about similarity? (large value $\Rightarrow$ similar, 0 or close to 0 $/Rightarrow$ very different)
Measures with a maximum value, e.g., Bray-Curtis, Morisita, Jaccard, Sorenson,
similarity = 1 - dissimilarity

Presence-absence data: $y_{ik} = 0$ or 1 for all species and sites
Usually written in terms of:
a: number of species only in sample i, i.e. in i but not both
b: number of species only in sample j, i.e. in j but not both
c: number of species in both samples

Sorenson:

$$\begin{aligned} d_{ij} &= \frac{a+b}{a+b+2c} = \frac{(a+b)/2}{(A+B)/2} \\ &= \frac{\mid y_{ik} - y_{jk} \mid}{Y_i + Y_j}, \text{ when } y\text{'s are 0 or 1.} \\ &\quad \text{(So Sorenson is BC on presence/absence data)} \end{aligned}$$

$A = a + c$ = total number species in i, $B = b + c$ = total number species in j.

Jaccard:

$$\begin{aligned} d_{ij} &= \frac{a+b}{a+b+c} \\ &= \frac{2S}{1+S} \end{aligned}$$

Continuous data:

Euclidean:

$$d_{ij} = \sqrt{\sum_k (y_{ik} - y_{jk})^2}$$

Manhattan:

$$d_{ij} = \sum_k \mid y_{ik} - y_{jk} \mid$$

Bray-Curtis:

$$d_{ij} = \frac{\sum_k \mid y_{ik} - y_{jk} \mid}{\sum_k (y_{ik} + y_{jk})} = \frac{\sum_k \mid y_{ik} - y_{jk} \mid}{Y_i + Y_j}$$

Canberra:

$$d_{ij} = \frac{1}{\#\text{ non-zero entries}} \sum_k \left[\frac{\mid y_{ik} - y_{jk} \mid}{y_{ik} + y_{jk}} \right]$$

Bray-Curtis on proportional abundance, first compute $p_{ik} = y_{ik}/Y_i$ for each sample :

$$\begin{aligned} d_{ij} &= \frac{\sum_k \mid p_{ik} - p_{jk} \mid}{\sum_k (p_{ik} + p_{jk})} \\ &= \frac{1}{2} \sum_k \mid p_{ik} - p_{jk} \mid \end{aligned}$$

Morisita-Horn, computed using y's, p's computed "internally":

$$\begin{aligned} d_{ij} &= 1 - \frac{2 \sum_k (y_{ik}/Y_i)(y_{jk}/Y_j)}{\sum_k (y_{ik}/Y_i)^2 + \sum_k (y_{jk}/Y_j)^2} \\ &= 1 - \frac{2 \sum_k p_{ik} p_{jk}}{\sum_k p_{ik}^2 + \sum_k p_{jk}^2} \\ &= \frac{\sum_k (p_{ik} - p_{jk})^2}{\sum_k p_{ik}^2 + \sum_k p_{jk}^2} \end{aligned}$$