# Distance measures

The distance measures defined by L&L and calculated by PRIMER 7 (in addition to $D_1$) are:

$ D_2 = \sqrt{ \frac{1}{p} \sum_i \left( y_{i1} - y_{i2} \right) ^ 2 } \text{ \hspace{25mm} average distance,} $

where the number of species *p* is fixed for all pairs of samples, so this is a constant multiple of Euclidean distance $D_1$ and will therefore give identical dendrograms, ordinations etc. (complete data is assumed for all these formulae, i.e. without missing values, though automatic adjustment to formulae under *pairwise elimination* of missing values is carried out for all measures, see later);

$ D_3 = \sqrt{ 2 \left( 1 - \frac{ \sum_i y_{i1} y_{i2} }{ \sqrt{ \sum_i y_{i1}^2 \sum_i y_{i2}^2} } \right) } \text{ \hspace{18mm} Orloci's chord distance;} $

$ D_4 = \text{arccos} \left(1 - \frac{1}{2} D_3^2 \right) \text{ \hspace{27mm} geodesic metric;} $

$ D_6 = \left( \sum_i \left| y_{i1} - y_{i2} \right| ^r \right) ^{1/r} \text{ \hspace{26mm} Minkowski metric,} $

where r can be specified by the user (note *r*=1 gives Manhattan, and *r*=2 Euclidean distance);

$ D_7 = \sum_i \left| y_{i1} - y_{i2} \right| \text{ \hspace{34mm} Manhattan distance, }$

whose use of absolute rather than squared differences confers slightly better robustness to outliers

$ D_8 = \frac{1}{p_{12}} \sum_i \left| y_{i1} - y_{i2} \right| \text{ \hspace{25mm} Czekanowski's mean character difference,} $

in the form where p12 is the number of species that are not jointly absent in samples 1 and 2 (the changing denominator across pairs of samples, from excluding joint absences, can make a big difference to a coefficient's behaviour, so is indicated clearly by 'exc0-0' in the drop-down box).

$ D_{10} = \sum_i \frac{ \left| y_{i1} - y_{i2} \right| }{ \left( y_{i1} + y_{i2} \right) } \text{ \hspace{33mm} Canberra metric of Lance \& Williams,} $

which must exclude joint absences so that it can be defined, but is less useful than its averaged form, divided by p12, found as Canberra similarity in the quantitative similarity list;

$ D_{11} = \sqrt{ \frac{1}{p_{12}} \sum_i \left( \frac{ y_{i1} - y_{i2} }{ y_{i1} + y_{i2} } \right)^2 } \text{ \hspace{22mm} Clark's coefficient of divergence,} $

also in the form in which double zeros are excluded from the summation and the divisor $p_{12}$;

$ D_{15} = \sqrt{ \sum_i \frac{1}{y_{i+}} \left( \frac{ y_{i1} }{\sum_i y_{i1}} - \frac{y_{i2}}{ \sum_i y_{i2} } \right)^2 } \text{ \hspace{15mm}} \chi^2 \text{(chi-squared) metric,} $

where $y_{i+} = \sum_j y_{ij}$, the sum across all samples of the entries for the $i$th species, and effectively the same, to within a constant, as the following;

$ D_{16} = \sqrt{ \sum_i \frac{1}{y_{i+}/ \sum_i y_{i+}} \left( \frac{ y_{i1} }{\sum_i y_{i1}} - \frac{y_{i2}}{ \sum_i y_{i2} } \right)^2 } \text{ \hspace{11mm}} \chi^2 \text{distance,} $

the implicit distance underlying Correspondence Analysis, which is seen to be a type of Euclidean distance, from samples which are standardised by their totals across species, and then inversely weighted by species totals across samples (the double standardisation being responsible for the practical difficulties $\chi^2$ distance can have with rare species, for which the divisor is near zero); and

$ D_{17} = \sqrt{ \sum_i \left( \sqrt{ \frac{ y_{i1} }{ \sum_i y_{i1}}} - \sqrt{ \frac{ y_{i2} }{ \sum_i y_{i2} } } \right)^2 } \text{ \hspace{5mm} Hellinger distance, advocated by Rao,}$

the only omission above being $D_{13}$, which is simply the complement of Sørensen similarity, $S_8$.

---

Revision #53
Created 5 June 2024 01:42:12 by Arden
Updated 15 January 2025 02:29:12 by Abby Miller