Binary divisive clustering

Two new clustering methods are introduced towards the end of Chapter 3 in CiMC, the first still a hierarchical clustering method leading to a tree diagram, but a divisive rather than agglomerative algorithm in which all samples start off in a single group and are then split into two groups, each of those then further sub-divided into two, and so on until some stopping rule is activated. The sub-groups are not constrained to be of comparable sizes, in fact may sometimes be a split of *n* samples into a group of size *n*–1 and a singleton. In keeping with the principles embodied by the PRIMER package, the criterion which is maximised in making each split is the non-parametric ANOSIM *R* statistic of Section 9, used as a pure measure of group separation for a multivariate set of samples (and not in any way as a test statistic). *R* is essentially the difference between the averages of rank dissimilarities between two groups and averaged rank dissimilarity within those groups, suitably scaled so that it takes values up to +1 (*perfect* rank separation, in which all dissimilarities between the groups are larger than any dissimilarities within either group). After each binary division, the dissimilarities among samples within each new group are re-ranked, and used to maximise *R* in a further binary division. Even for quite modest sample sizes, evaluating *R* for all possible splits into two groups can be prohibitive, so a search algorithm is required and the number of random restarts of that process needs to be specified (default 10, but increase this if the routine runs quickly). A range of different stopping rules are allowed, which can be used in combination: a) a split which would produce a group of size *n* or less is never made (*n* specified); b) groups of size <*n* are never split (*n* specified); c) a split is not made if the largest *R* is less than a specified value; d) a group is never split if a SIMPROF test of its samples cannot reject the hypothesis of ‘no structure’ within that group – this is the least arbitrary and most natural of the stopping rules, a natural counterpart to the stopping rule for interpretation used for the agglomerative clustering described earlier.

A parallel routine **Analyse>Cluster>LINKTREE** is described in Section 13 (called *linkage trees*), a constrained divisive clustering in which binary splits of, for example, biotic community samples are made in the same way (by maximising *R*), but only if an environmental variable can be found that takes a non-overlapping range of values in the two groups produced (a possible ‘explanation’ for that split therefore). In contrast, this new routine to PRIMER 7 is a completely *unconstrained tree*, accessed by **Analyse>Cluster>UNCTREE**: each sample is divided to maximise *R*, based only on the input resemblance matrix, e.g. the community similarities, without external constraints.

Revision #1
Created 19 June 2024 23:08:31 by Arden
Updated 19 June 2024 23:15:49 by Arden