3.7 k-R clustering (non-hierarchical)

Another major class of clustering techniques is non-hierarchical, referred to above as *flat clustering.* The desired number of clusters (*k*) must be specified in advance, and an iterative search attempts to divide the samples in an optimal way into *k* groups, in one operation rather than incrementally. The classic method, the idea of which was outlined in the two-group case above, is *k-means clustering*, which seeks to minimise within-group sums of squares about the *k* group centroids. This is equivalent to minimising some weighted combination of within-group resemblances between pairs of samples, as measured by a squared Euclidean distance coefficient (you can visualise this by adding additional groups to Fig. 3.8). The idea can again be generalised to apply to *any* resemblance measure, e.g. Bray-Curtis, by maximising ANOSIM *R*, which measures (non-parametrically) the degree of overall separation of the *k* groups, formed from the ranks in the full resemblance matrix. (Note that we defined equation (3.1) as if it applied only to two groups, but the definition of *R* is exactly the same for the *k*-group case, equation (6.1)). By analogy with *k*-means clustering, the principle of maximising *R* to obtain a *k*-group division of the samples is referred to as *k-R clustering*, and it will again involve an iterative search, from several different random starting allocations of samples to the *k* groups.

Experience with *k*-means clustering suggests that a flat clustering of the *k-R* type should sometimes have slight advantages over a hierarchical (agglomerative or divisive) method, since samples are able to move between different groups during the iterative process. The *k*-group solution will not, of course, simply split one of the groups in the (*k*-1)-group solution: there could be a widescale rearrangement of many of the points into different groups. A widely perceived disadvantage of the *k*-means idea is the need to specify *k* before entering the routine, or if it is re-run for many different *k* values, the absence of a convenient visualisation of the clustering structure for differing values of *k*, analogous to a dendrogram. This has tended to restrict its use to cases where there is a clear *a priori* idea of the approximate number of groups required, perhaps for operational reasons (e.g. in a quality classification system). However, the SIMPROF test can also come to the rescue here, to provide a choice of *k* which is objective. Starting from a low value for *k* (say 2) the two groups produced by *k-R* clustering are tested for evidence of within-group structure by SIMPROF. If either of the tests are significant, the routine increments *k* (to 3), finds the 3-group solution and retests those groups by SIMPROF. The procedure is repeated until a value for *k* is reached in which none of the *k* groups generates significance in their SIMPROF test, and the process terminates with that group structure regarded as the best solution. (This will not, in general, correspond to the maximum *R* when these optima for each *k* are compared across all possible *k*; e.g. R must increase to its maximum of 1 as *k* approaches *n*, the number of samples.)

Fig. 3.10c shows the optimum grouping produced by *k-R* clustering, superimposed on the same MDS plot as for Figs. 3.10 a & b. The SIMPROF routine has again terminated the procedure with *k* =4 groups (A to D), which are very similar to those for the two hierarchical methods, but with the three sites 9, 23 and 24 allocated to the four groups in yet a third way. This appears to be at least as convincing an allocation as for either of the hierarchical plots (though do not lose sight of the fact that the MDS itself is only an approximation to the real inter-sample resemblances).

Average rank form of flat clustering

A variation of this flat-clustering procedure does not use *R* but a closely related statistic, arising from the concept of group-average linking met earlier in Table 3.2. For a pre-specified number of groups (*k*), every stage of the iterative process involves removing each sample in turn and then allocating it to one of the *k*-1 other groups currently defined, or returning it to its original group. In *k*-*R* clustering it is re-allocated to the group yielding the highest *R* for the resulting full set of groups. In the *group average rank* variation, the sample is re-allocated to the group with which it has greatest (rank) similarity, defined as the average of the pairwise values (from the ranked form of the original similarity matrix) between it and all members of that group – or all of the remaining members, in the case of its original group. The process is iterated until it converges and repeated a fair number of times from different random starting allocations to groups, as before. The choice of *k* uses the same SIMPROF procedure as previously, and it is interesting to note that, for the Bristol Channel zooplankton data, this group-average variation of *k*-*R* clustering produces exactly the same four groups as seen in Fig 3.10c. This will not always be the case because the statistic here is subtly different than the ANOSIM *R* statistic, but both exploit the same non-parametric form of the resemblance matrix so it should be expected that the two variations will give closer solutions to each other than to the hierarchical methods.

In conclusion

A ‘take-home’ message from Fig. 3.10 is that clustering rarely escapes a degree of arbitrariness: the data simply may not represent clearly separated clusters. For the Bristol Channel sites, where there certainly are plausible groups but within a more or less continuous gradation of change in plankton communities (strongly correlated with increased average salinity of the sites), different methods must be expected to chop this continuum up in slightly different ways. Use of a specific grouping from an agglomerative hierarchy should probably be viewed operationally as little worse (or better) than that from a divisive hierarchy or from the non-hierarchical k-R clustering, in either form; it is reassuring here that SIMPROF supports four very similar groups for all these methods. In fact, especially in cases where a low-dimensional MDS plot is not at all reliable because of high stress (see Chapter 5), the plurality of clustering methods does provide some insight into the robustness of conclusions that can be drawn about group structures from the ‘high-dimensional’ resemblance matrix. Such comparisons of differing clustering methods need to ‘start from the same place’, namely using the same resemblance matrix, otherwise an inferred lack of a stable group structure could be due to the differing assumptions being made about how the (dis)similarity between two samples is defined (e.g. Bray-Curtis *vs* squared Euclidean distance). This is also a point to bear in mind in the following chapters on competing ordination methods: a primary difference between them is often not the way they choose to represent high-dimensional information in lower dimensional space but how they define that higher-dimensional information differently, in their choice of explicit or implicit resemblance measure.

Revision #7
Created 20 February 2022 09:10:53 by Arden
Updated 27 June 2023 23:40:09 by Arden