Journal of Experimental Marine Biology and Ecology 483 (2016) 147–155

Contents lists available at ScienceDirect

Journal of Experimental Marine Biology and Ecology

journal homepage: www.elsevier.com/locate/jembe

Clustering in non-parametric multivariate analyses

K. Robert Clarke [a,b,*], Paul J. Somerfield [a], Raymond N. Gorley [b]

[a] Plymouth Marine Laboratory, Prospect Place, West Hoe, Plymouth PL1 3DH, UK
[b] Primer-E Ltd, c/o Plymouth Marine Laboratory, Prospect Place, West Hoe, Plymouth PL1 3DH, UK

ARTICLE INFO

Article history:
Received 16 June 2016
Received in revised form 13 July 2016
Accepted 15 July 2016
Available online 22 July 2016

Keywords:
Non-parametric multivariate
Divisive clustering
Flat clustering
SIMPROF
Cophenetic correlation
Cophenetic distance

ABSTRACT

Non-parametric multivariate analyses of complex ecological datasets are widely used. Following appropriate pre-treatment of the data inter-sample resemblances are calculated using appropriate measures. Ordination and clustering derived from these resemblances are used to visualise relationships among samples (or variables). Hierarchical agglomerative clustering with group-average (UPGMA) linkage is often the clustering method chosen. Using an example dataset of zooplankton densities from the Bristol Channel and Severn Estuary, UK, a range of existing and new clustering methods are applied and the results compared. Although the examples focus on analysis of samples, the methods may also be applied to species analysis. Dendrograms derived by hierarchical clustering are compared using cophenetic correlations, which are also used to determine optimum β in flexible beta clustering. A plot of cophenetic correlation against original dissimilarities reveals that a tree may be a poor representation of the full multivariate information. UNCTREE is an unconstrained binary divisive clustering algorithm in which values of the ANOSIM *R* statistic are used to determine (binary) splits in the data, to form a dendrogram. A form of flat clustering, *k-R* clustering, uses a combination of ANOSIM *R* and Similarity Profiles (SIMPROF) analyses to determine the optimum value of *k*, the number of groups into which samples should be clustered, and the sample membership of the groups. Robust outcomes from the application of such a range of differing techniques to the same resemblance matrix, as here, result in greater confidence in the validity of a clustering approach.

1. Introduction

Clustering is a numerical process that groups a set of objects (samples, variables) so that objects in the same group, or cluster, are in some sense more similar to each other than they are to objects in another cluster. Clustering is widely used in scientific investigations ranging from numerical taxonomy and bioinformatics to image analysis and data processing. In ecological studies clustering is often applied explicitly to data, to determine which samples or variables cluster together. In many ecological analyses, however, a classification step is implicit. Examples include clustering nucleic acid sequences into groups (Blaxter et al., 2005) to define operational taxonomic units (OTUs), or developing hierarchical trees based on taxonomic relationships (Faith, 1992) or ecological traits (Petchey and Gaston, 2002) in order to calculate some index of community diversity. Although widely used in some branches of ecology, and central to the methodology in others, little consideration is generally given to the range of clustering options potentially available, and the consequences of choosing one approach over another. Literally hundreds of clustering methods exist, some of them operating on resemblance matrices while others are based on the original data (Legendre and Legendre, 2012). Everitt (1980) and Cormack (1971) give excellent and readable reviews, while Clifford and Stephenson (1975) is another well-established text from an ecological viewpoint. To cope with this variety, in ecological studies, a widely adopted approach has been to use a single technique that has been found to be of widespread utility, while recommending the need to perform a cluster analysis in conjunction with a range of other techniques (e.g. ordination, statistical testing) to obtain balanced and reliable conclusions (Clarke et al., 2014).

Hierarchical clustering with group-average linking, based on sample similarities or dissimilarities such as the Bray-Curtis coefficient, has proved a useful technique in many ecological studies over the past half-century. As with clustering methods in general, it is appropriate for delineating groups of sites with distinct community structure. It is an agglomerative method, meaning that all samples start as singleton clusters, and the process proceeds by successively merging (agglomerating) pairs of clusters until all clusters have been merged into a single group. For this reason it is often termed hierarchical agglomerative clustering (HAC). Agglomerative methods are bottom-up and 'see' only the nearby points throughout much of the process. When reaching the top of the dendrogram no possibility of taking a different view, of the main merged groups that have formed, remains. Binary divisive methods, however, are potentially advantageous for some clustering

* Corresponding author at: Plymouth Marine Laboratory, Prospect Place, West Hoe, Plymouth PL1 3DH, UK.
E-mail addresses: krc@pml.ac.uk (K.R. Clarke), pjso@pml.ac.uk (P.J. Somerfield), tech@primer-e.com (R.N. Gorley).

http://dx.doi.org/10.1016/j.jembe.2016.07.010
0022-0981/

situations. They take a top-down view of the samples, so that the initial binary splits should (in theory) be better able to respect any major groupings in the data, since these are found first. However, as with all hierarchical methods, once a sample has been placed within one initial group it cannot jump to another at a later stage. While divisive methods have the potential to produce marginally better solutions in practice, there is a counterbalancing downside to their algorithms, in that they can be computationally intensive and complex (Gower, 1967), so iterative approaches are generally required. The agglomerative approach, in contrast, is simple and entirely determined, requiring nothing more than simple numerical operations based on values of resemblance measures.

Field et al. (1982) described a robust non-parametric multivariate strategy for the analysis of biological assemblage data, such as the abundance or biomass of taxa in samples. Collins and Williams (1982) present one of the first applications of the strategy, to plankton data from the Bristol Channel and Severn Estuary. In essence the strategy, expanded and clarified by Clarke (1993), is to display patterns among samples determined by appropriate resemblance measures (Clarke et al., 2006) using clustering and ordination, and to analyse these patterns using a range of hypothesis tests and associated analyses, primarily based on ranked resemblances. Additional analyses are constantly added to the framework. Clarke et al. (2008) described a method for divisive clustering constrained by thresholds in explanatory variables, Linkage Trees, and Similarity Profiles analysis (SIMPROF) which tests for multivariate structure within groups of samples. The latter was further developed by Somerfield and Clarke (2013) in the context of species (r-mode) analysis. The purpose of this paper is to compare and discuss methods for, and associated with, cluster analysis in this non-parametric multivariate framework. Some existing methods are considered in terms of how their success, in conserving the inter-sample patterns in the underlying resemblance matrix, may be assessed and compared. New clustering methods are introduced and their results are compared. For the purpose of the paper only examples based on analyses among samples are used, though it should be remembered that clustering of variables (taxa, functional groups, OTUs, environmental measurements) is often entirely appropriate following suitable pre-treatment of the data (Somerfield and Clarke, 2013).

2. Material and methods

2.1. Hierarchical agglomerative clustering

The most commonly used clustering techniques are hierarchical agglomerative methods. These usually take a resemblance matrix (Clarke et al., 2006) as their starting point and successively fuse the samples into groups, and the groups into larger clusters, starting with the highest mutual similarities then lowering the similarity level at which groups are formed, ending when all samples are in a single cluster. The result of a hierarchical clustering is generally presented as a tree diagram or dendrogram. There is no firm convention for which way up a dendrogram should be portrayed (increasing or decreasing resemblance values) or even whether the tree can be placed on its side, but we will refer to the *x*-axis as representing the full set of samples and the *y*-axis defining a resemblance level at which two samples or groups are considered to have fused. Neither is there anything sacrosanct about the ordering of samples along the *x*-axis, with the exception of constraints imposed by the grouping structure among samples at higher levels in the tree.

2.2. Linkage options

Within hierarchical agglomerative clustering several linkage/sorting/joining options are defined which determine how resemblances between samples and groups of samples are recalculated following fusion of samples into a group. For single linkage (also called nearest-neighbour joining) the dissimilarity of groups A and B, $\delta_{A\text{-}B}$, is the minimum across all dissimilarities between pairs of samples with the first in A and the second in B. The dissimilarity of a group C to two merged groups A and B, $\delta_{C\text{-}AB}$, is therefore just the minimum of $\delta_{C\text{-}A}$ and $\delta_{C\text{-}B}$. For complete linkage (also called farthest-neighbour joining), $\delta_{C\text{-}AB}$ is the maximum of $\delta_{C\text{-}A}$ and $\delta_{C\text{-}B}$. In group-average linkage $\delta_{A\text{-}B}$ is the simple (unweighted) average over all dissimilarities from A to B pairs, leading to the acronym UPGMA, Unweighted Pair Group Method with Arithmetic mean. When A and B are of different sizes, it follows that, under UPGMA, $\delta_{C\text{-}AB}$ is a weighted average of $\delta_{C\text{-}A}$ and $\delta_{C\text{-}B}$, e.g. giving more weight to $\delta_{C\text{-}A}$ if there are more samples in A than B. Somewhat confusingly, the simple average of $\delta_{C\text{-}A}$ and $\delta_{C\text{-}B}$ is then referred to as weighted linkage, WPGMA, since it weights the original dissimilarities between samples in C and those in the combined group A and B unequally.

Other linkage options have been suggested. One is the flexible beta method of Lance and Williams (1967), in which $\delta_{C\text{-}AB} = (1 - \beta)[(\delta_{C\text{-}A} + \delta_{C\text{-}B}) / 2] + \beta\delta_{A\text{-}B}$. Only negative values of β in the range $(-1, 0)$ make much sense in theory, the effect of including the $\delta_{A\text{-}B}$ term then being to make the merged AB group more likely to join with the group C, the further A and B themselves are from each other. That is, there will be a tendency to merge loosely bound samples or groups with each other, leaving tightly bound groups separate. Lance and Williams (1967) suggest the use of $\beta = -0.25$, for which the flexible beta has affinities with Gower's median method (Gower, 1967). If $\beta = 0$, $\delta_{C\text{-}AB} = (\delta_{C\text{-}A} + \delta_{C\text{-}B}) / 2$, which is the WPMGA method given above, also known as McQuitty's (1967) linkage.

Within a non-parametric multivariate analytical framework it might be expected that a linkage option that is a function only of the ranks in the underlying resemblance matrix would be preferred. Single linkage does this, but experience shows that it leads to 'chaining' in the resulting dendrogram, with samples continuously joined to the next most similar sample without forming discrete clusters. Complete linkage, conversely, tends to result in starkly separated, compact clusters. Group average linkage will find a seemingly reasonable balance between the two. In order to choose between linkage methods and their associated dendrograms a more objective means than simple visual comparison of dendrograms is clearly needed.

2.3. Cophenetic correlation

One objective approach is provided by cophenetic correlation, which is a (Pearson) matrix correlation between each original dissimilarity and the (vertical) distance through a dendrogram to the common node of the corresponding pair of samples (Jain and Dubes, 1988). If the *y*-axis of the dendrogram is a dissimilarity scale then, naturally, these vertical distances are also dissimilarities. A dendrogram is a good representation of the dissimilarity matrix, therefore, if the cophenetic correlation is close to 1. As such the correlation may be seen as a way to compare different dendrograms, to assess the performance of different analysis choices starting from the same dissimilarity matrix. In particular, the correlation may also be used to determine β for the flexible beta method, computing a range of values and choosing that which maximises the cophenetic correlation.

2.4. Binary divisive clustering

In hierarchical agglomerative clustering, samples start in separate groups and are successively merged until, at some level of similarity, all are considered to belong to a single group. Hierarchical divisive clustering does the converse operation: samples start in a single group and are divided into two sub-groups, which may be of quite unequal size, each of those being further sub-divided into two (i.e. binary division), and so on. Ultimately, all samples become singleton groups unless (preferably) some criterion is applied to stop further sub-division of any specific group.