Multivariate Analysis of Ecological Data

MICHAEL GREENACRE
Professor of Statistics at the Pompeu Fabra University in Barcelona, Spain

RAUL PRIMICERIO
Associate Professor of Ecology, Evolutionary Biology and Epidemiology
at the University of Tromsø, Norway

Chapter 7 Offprint

Hierarchical Cluster Analysis

First published: December 2013
ISBN: 978-84-92937-50-9

Supporting websites:
www.fbbva.es
www.multivariatestatistics.org

Fundación BBVA

Chapter 7

Hierarchical Cluster Analysis

In Part 2 (Chapters 4 to 6) we defined several different ways of measuring distance (or dissimilarity as the case may be) between the rows or between the columns of the data matrix, depending on the measurement scale of the observations. As we remarked before, this process often generates tables of distances with even more numbers than the original data, but we will now show how this step actually simplifies our understanding of the data. Distances between objects can be visualized in many simple and evocative ways. In this chapter we shall consider a graphical representation of a matrix of distances which is perhaps the easiest to understand – a dendrogram, or tree – where the objects are joined together in a hierarchical fashion from the closest, that is most similar, to the furthest apart, that is the most different. The method of hierarchical cluster analysis is best explained by describing the algorithm, or set of instructions, which creates the dendrogram result. In this chapter we demonstrate the application of hierarchical clustering on a small example and then list the different variants of the method that are possible.

Contents

The algorithm for hierarchical clustering

As an example we shall consider again the small data set in Exhibit 5.6: seven samples on which 10 species are indicated as being present or absent. In Chapter 5 we discussed two of the many dissimilarity coefficients that are possible to define between the samples: the first based on the matching coefficient and the second based on the Jaccard index. The latter index counts the number of "mismatches" between two samples after eliminating the species that do not occur in either of the pair. Exhibit 7.1 shows the complete table of inter-sample dissimilarities based on the Jaccard index.

Exhibit 7.1: *Dissimilarities, based on the Jaccard index, between all pairs of seven samples in Exhibit 5.6. Both the lower and upper triangles of this symmetric dissimilarity matrix are shown here (the lower triangle is outlined as in previous tables of this type)*

Samples	A	B	C	D	E	F	G
A	0.000	0.500	0.429	1.000	0.250	0.625	0.375
B	0.500	0.000	0.714	0.833	0.667	0.200	0.778
C	0.429	0.714	0.000	1.000	0.429	0.667	0.333
D	1.000	0.833	1.000	0.000	1.000	0.800	0.857
E	0.250	0.667	0.429	1.000	0.000	0.778	0.375
F	0.625	0.200	0.667	0.800	0.778	0.000	0.750
G	0.375	0.778	0.333	0.857	0.375	0.750	0.000

The first step in the hierarchical clustering process is to look for the pair of samples that are the most similar, that is the closest in the sense of having the lowest dissimilarity – this is the pair B and F, with dissimilarity equal to 0.2.[1] These two samples are then joined at a level of 0.2 in the first step of the dendrogram, or clustering tree (see the first diagram of Exhibit 7.3, and the vertical scale of 0 to 1 which calibrates the level of clustering). The point at which they are joined is called a *node*.

We are basically going to keep repeating this step, but the only problem is how to calculate the dissimilarity between the merged pair (B,F) and the other samples. This decision determines what type of hierarchical clustering we intend to perform, and there are several choices. For the moment, we choose one of the most popular ones, where the dissimilarity between the merged pair and the others will be the maximum of the pair of dissimilarities in each case. For example, the dissimilarity between B and A is 0.500, while the dissimilarity between F and A is 0.625. Hence we choose the maximum of the two, 0.625, to quantify the dissimilarity between (B,F) and A. Continuing in this way we obtain a new dissimilarity matrix Exhibit 7.2.

The process is now repeated: find the smallest dissimilarity in Exhibit 7.2, which is 0.250 for samples A and E, and then cluster these at a level of 0.25, as shown in the second figure of Exhibit 7.3. Then recompute the dissimilarities between the merged pair (A,E) and the rest to obtain Exhibit 7.4. For example, the dissimilarity between (A,E) and (B,F) is the maximum of 0.625 (A to (B,F)) and 0.778 (E to (B,F)).

[1] Recall what this value means: five species occurred in at least one of the samples B and F, four occurred in both, while one was present in B but not in F, so the Jaccard index of dissimilarity is 1/5 = 0.2.

Samples	A	(B,F)	C	D	E	G
A	0.000	0.625	0.429	1.000	0.250	0.375
(B,F)	0.625	0.000	0.714	0.833	0.778	0.778
C	0.429	0.714	0.000	1.000	0.429	0.333
D	1.000	0.833	1.000	0.000	1.000	0.857
E	0.250	0.778	0.429	1.000	0.000	0.375
G	0.375	0.778	0.333	0.857	0.375	0.000

Exhibit 7.2: *Dissimilarities calculated after* B *and* F *are merged, using the "maximum" method to recompute the values in the row and column labelled* (B,F)

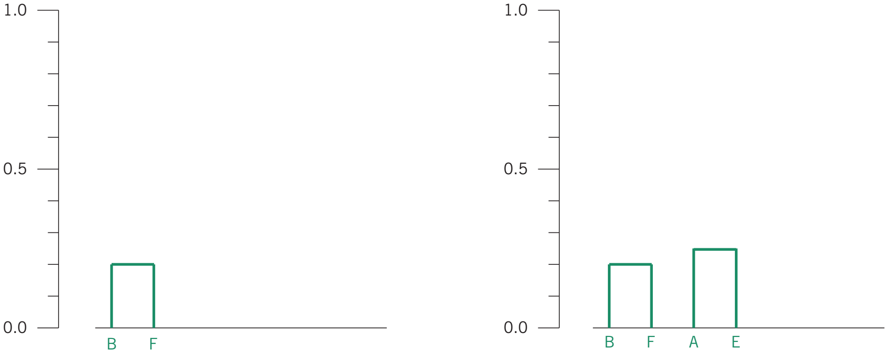

Exhibit 7.3: *First two steps of hierarchical clustering of Exhibit 7.1, using the "maximum" (or "complete linkage") method*

Samples	(A,E)	(B,F)	C	D	G
(A,E)	0.000	0.778	0.429	1.000	0.375
(B,F)	0.778	0.000	0.714	0.833	0.778
C	0.429	0.714	0.000	1.000	0.333
D	1.000	0.833	1.000	0.000	0.857
G	0.375	0.778	0.333	0.857	0.000

Exhibit 7.4: *Dissimilarities calculated after* A *and* E *are merged, using the "maximum" method to recompute the values in the row and column labelled* (A,E)

In the next step the lowest dissimilarity in Exhibit 7.4 is 0.333, for C and G – these are merged, as shown in the first diagram of Exhibit 7.6, to obtain Exhibit 7.5. Now the smallest dissimilarity is 0.429, between the pair (A,E) and (B,G), and they are shown merged in the second diagram of Exhibit 7.6. Exhibit 7.7 shows the last two dissimilarity matrices in this process, and Exhibit 7.8 the final two steps of the construction of the dendrogram, also called a *binary tree* because at each step two objects (or clusters of objects) are merged. Because there are 7 objects to be clustered, 6 nodes are formed in the sequential process (i.e., one

Exhibit 7.5: *Dissimilarities calculated after* C *and* G *are merged, using the "maximum" method to recompute the values in the row and column labelled* (C,G)

Samples	(A,E)	(B,F)	(C,G)	D
(A,E)	0.000	0.778	0.429	1.000
(B,F)	0.778	0.000	0.778	0.833
(C,G)	0.429	0.778	0.000	1.000
D	1.000	0.833	1.000	0.000

Exhibit 7.6: *The third and fourth steps of hierarchical clustering of Exhibit 7.1, using the "maximum" (or "complete linkage") method. The point at which objects (or clusters of objects) are joined is called a node*

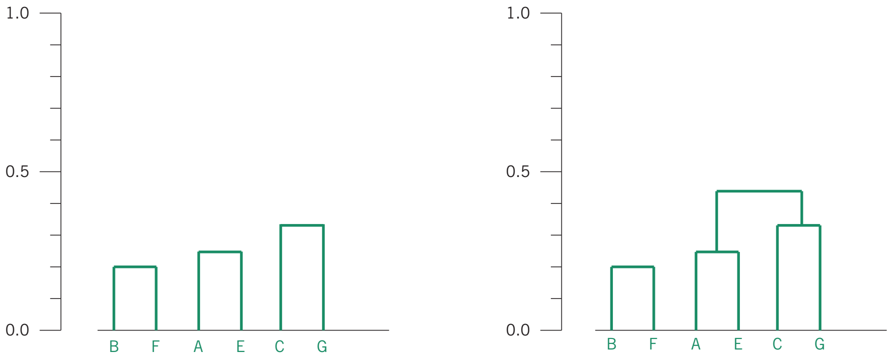

Exhibit 7.7: *Dissimilarities calculated after* C *and* G *are merged, using the "maximum" method to recompute the values in the row and column labelled* (C,G)

Samples	(A,E,C,G)	(B,F)	D
(A,E,C,G)	0.000	0.778	1.000
(B,F)	0.778	0.000	0.833
D	1.000	0.833	0.000

Samples	(A,E,C,G,B,F)	D
(A,E,C,G,B,F)	0.000	1.000
D	1.000	0.000

less than the number of objects) to arrive at the final tree where all objects are in a single cluster.

Cutting the tree

The final dendrogram on the right of Exhibit 7.8 is a compact visualization of the dissimilarity matrix in Exhibit 7.1, computed on the presence-absence data of Exhibit 5.6. Interpretation of the structure of data is made much easier now – we can see that there are three pairs of samples that are fairly close, two of these pairs [(A,E) and (C,G)] are in turn close to each other, while the sin-

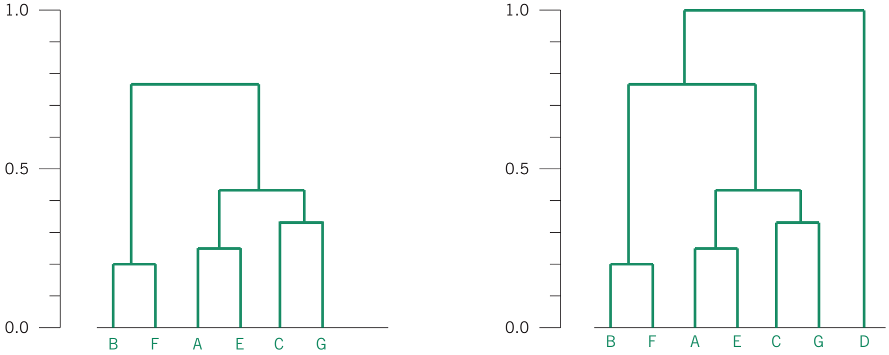

Exhibit 7.8:
The fifth and sixth steps of hierarchical clustering of Exhibit 7.1, using the "maximum" (or "complete linkage") method. The dendrogram on the right is the final result of the cluster analysis

gle sample D separates itself entirely from all the others. Because we used the "maximum" method, all samples clustered below a particular level of dissimilarity will have inter-sample dissimilarities less than that level. For example, 0.5 is the point at which samples are exactly as similar to one another as they are dissimilar, so if we look at the clusters of samples below 0.5 – i.e., (B,F), (A,E,C,G) and (D) – then within each cluster the samples have more than 50% similarity, in other words more than 50% co-presences of species. The level of 0.5 also happens to coincide in the final dendrogram with a large jump in the clustering levels: the node where (A,E) and (C,G) are clustered is at level of 0.429, while the next node where (B,F) is merged is at a level of 0.778. This is thus a very convenient level to *cut* the tree to define clusters. If the branches are cut at 0.5, we are left with the three clusters of samples (B,F), (A,E,C,G) and (D), which can be labelled types 1, 2 and 3 respectively. In other words, we have created a categorical variable, with three categories, and the samples are classified as follows:

A	B	C	D	E	F	G
2	1	2	3	2	1	2

Checking back to Chapter 2, this is exactly the objective which we described in the lower right hand corner of the multivariate analysis scheme (Exhibit 2.2) – to reveal a categorical latent variable which underlies the structure of a data set.

Maximum, minimum and average clustering

Two crucial choices are necessary when deciding on a cluster analysis algorithm. The first is to decide how to quantify dissimilarities between two clusters: in the above illustration the Jaccard index was used. The second choice is how to update the matrix of dissimilarities at each step of the clustering:

in the algorithm described above the maximum value of the between-cluster dissimilarities was chosen. This is called the *maximum* method, also known as *complete linkage* cluster analysis, because a cluster is formed when all the dissimilarities ("links") between pairs of objects in the cluster are less then a particular level. There are several alternatives to complete linkage as a clustering criterion, and we only discuss two of these: minimum and average clustering.

The *minimum* method goes to the other extreme and forms a cluster when only one pair of dissimilarities (not all) is less than a particular level – this is known as *single linkage* cluster analysis. So at every updating step we choose the minimum of the two distances and two clusters of objects can be merged when there is a single close link between them, irrespective of the other inter-object distances. In general, this is not a suitable choice for most applications, because it can lead to clusters that are quite heterogeneous internally, and the usual object of clustering is to obtain homogeneous clusters.

The *average* method is an attractive compromise where at each step the dissimilarity between two clusters is the average of all the pairwise dissimilarities between the clusters. This is also dubbed UPGMA clustering in the literature, a rather laborious abbreviation for "Unweighted Pair-Group Method using Averages". Notice that this is not merely taking the arithmetic average of the two dissimilarity values available at each step for the updating, but rather taking into account the size of each cluster as well. For example, if one cluster contains two cases and another three, then there are six dissimilarities on which the (unweighted) arithmetic average needs to be computed – this is equivalent to weighting the clusters by their sample sizes in the updating of the average dissimilarity.

Validity of the clusters

If a cluster analysis is performed on a data matrix, a set of clusters can always be obtained, even if there is no actual grouping of the objects, in this case the samples. So how can we evaluate whether the three clusters in this example are not just any three groups which we might have obtained on random data with no underlying structure? We shall consider this question more closely in Chapter 17 when we deal with statistical inference and give one possible answer to this problem using a permutation test. Apart from this statistical issue there is also the substantive issue of where to cut the tree. In this example, the three clusters established by complete linkage were such that within each cluster all inter-sample dissimilarities were all less than 0.5. It would be difficult to justify cutting the tree at a higher level, because that would mean that some pairs of samples in a cluster would be more dissimilar than similar. But this substantive cut-off level of 0.5 is particular to the Jaccard index and to measures like the Bray-Curtis that have scales with a clear interpretation. If one uses an Euclidean distance, or chi-square distance, for example, their scales are not clearly interpretable and we have to resort to deciding on the number of clusters by an inspection of the tree, cutting the branches where there is a big jump in the level of successive nodes, or by a statistical criterion described in Chapter 17.

Clustering correlations on variables

Just like we clustered samples, so we can cluster variables in terms of their correlations. In this case it may be more intuitive to show cluster levels as the correlation, the measure of similarity, rather than the reverse measure of dissimilarity. The similarity based on the Jaccard index can also be used to measure association between species – the index counts the number of samples that have both species of the pair, relative to the number of samples that have at least one of the pair. Exhibit 7.9 shows the cluster analyses based on these two alternatives, for the columns of Exhibit 5.6. There are two differences here compared to previous dendrograms: first, the vertical scale is descending as the tree is being constructed from the bottom up, since the clustering is on similarities, and second, the R function `hclust` used to perform the clustering places the species labels at a constant distance below their initial clustering levels. The fact that these two trees are so different is no surprise: the first one based on the correlation coefficient takes into account the co-absences, which strengthens the correlation, while the second does not. Both have the pairs (*sp2*,*sp5*) and (*sp3*,*sp8*) at maximum similarity of 1 because these are identically present and absent across the samples. Species *sp1* and *sp7* are similar in terms of correlation, due to co-absences – *sp7* only occurs in one sample, sample E, which also has *sp1*, a species which is absent in four other samples. Notice in Exhibit 7.9(b) how species *sp10* and *sp1* both join the cluster (*sp2*,*sp5*) at the same level (0.5).

Clustering a large data set

The more objects there are to cluster, the more complex becomes the result, and we would not generally apply this method to a set of more than 100 objects, say. In Exhibit 4.5 we showed part of the matrix of standardized Euclidean distances between the 30 sites of Exhibit 1.1, and Exhibit 7.10 shows the hierarchical clustering of this distance matrix, using compete linkage. There are two obvious places where we can cut the tree, at about level 3.4, which gives four clusters, or about 2.7, which gives six clusters. Which one we should choose depends on substantive as well as statistical grounds. For example, the six-cluster solution splits a large group on the right hand side of the dendrogram into two; if this is usefully interpreted as two different sets of sites in the context of the study, then the six-cluster solution would be preferred. But there is also the statistical issue about whether that split can be considered random or not, which is what we will deal with in Chapter 17.

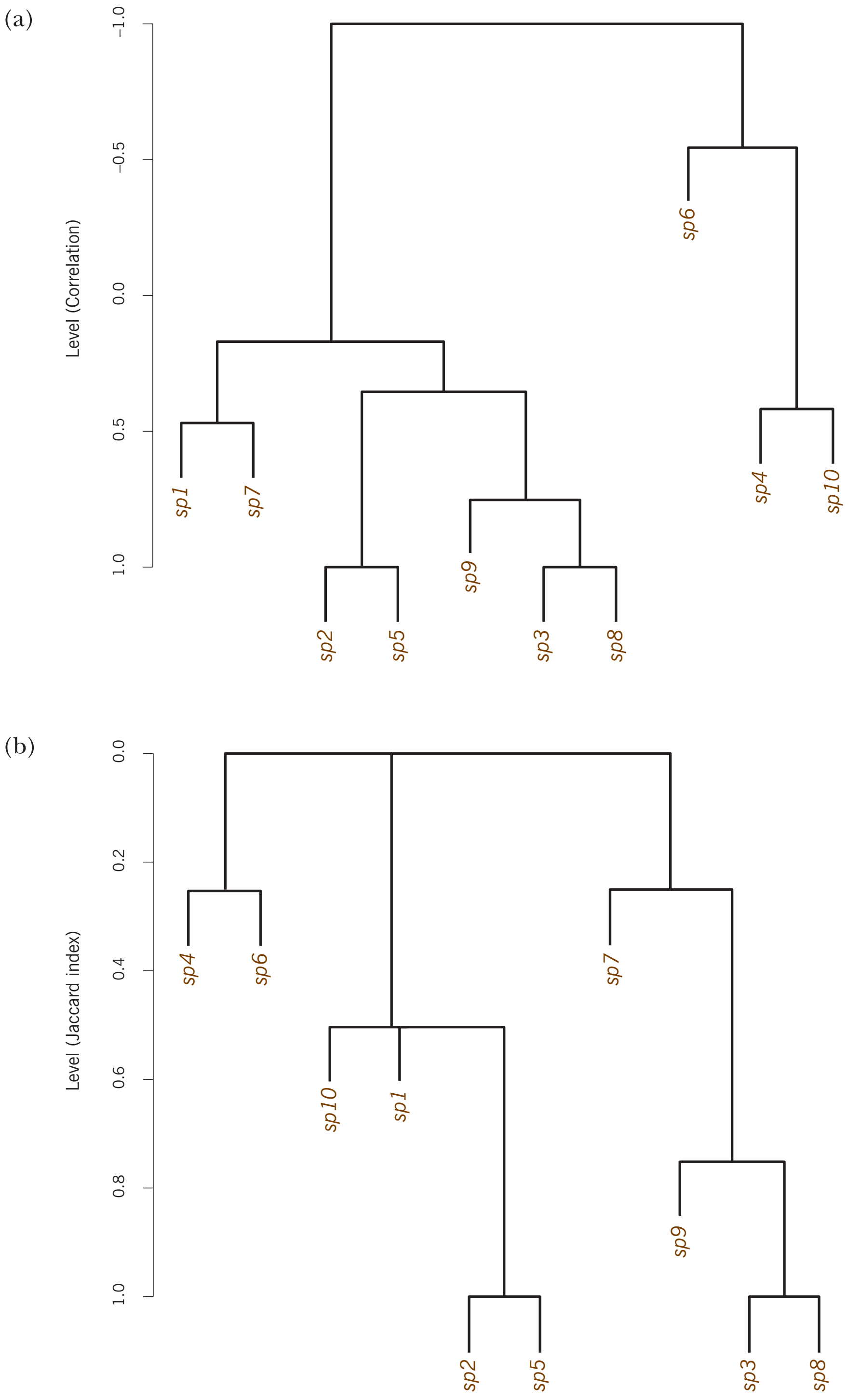

Exhibit 7.9: *Complete linkage cluster analyses of similarities between species: (a) r, the correlation coefficient between species); (b) Jaccard similarity index between species. The* R *function* `hclust` *which calculates the dendrograms places the object (species) labels at a constant distance below its clustering level*

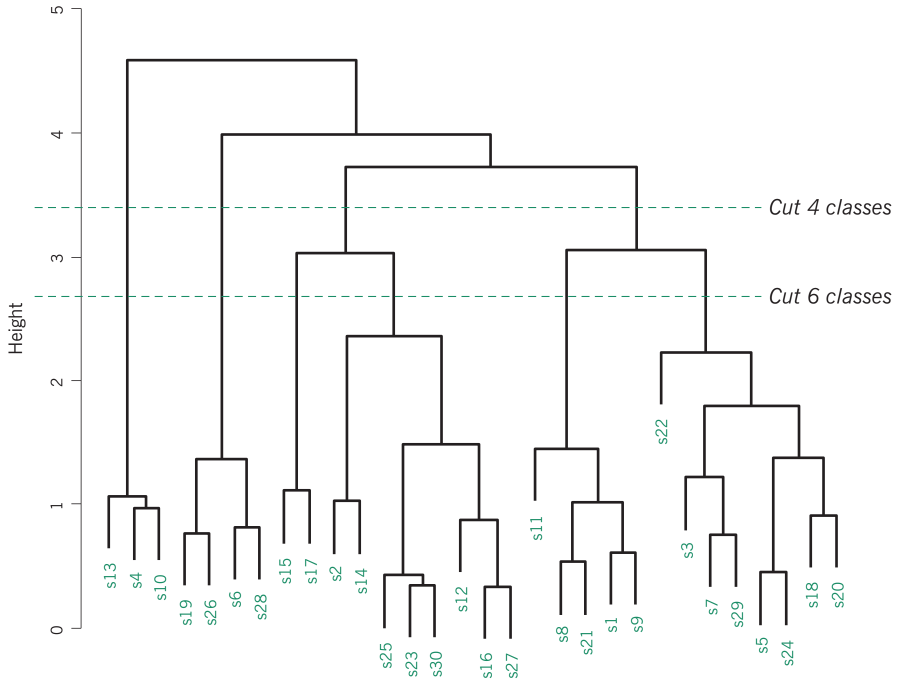

Exhibit 7.10:
Complete linkage cluster analyses of the standardized Euclidean distances of Exhibit 4.5

If the data set consists of a very large set of objects, say in the thousands, then nonhierarchical clustering can be used, as described in the next chapter, but the number of clusters desired has to be specified in advance. A hybrid approach could be to initially reduce (using nonhierarchical clustering) the very large set into a large number of very compact clusters, for example reducing a set of 1,000 objects to about 100 clusters with an average of 10 objects each, and then performing hierarchical clustering on the 100 clusters.

SUMMARY:
Hierarchical cluster analysis

1. Hierarchical cluster analysis of n objects is defined by a stepwise algorithm performed on a matrix of appropriately chosen dissimilarities or distances previously computed between the objects. Two objects are merged at each step, the two which have the least dissimilarity or distance.

2. As the algorithm proceeds, objects become clusters of objects, so we need to decide how to measure dissimilarity/distance between clusters. Some standard options are the maximum dissimilarity (complete linkage) between the objects of each cluster, the minimum dissimilarity (single linkage) or the average dissimilarity (average linkage).

3. The results of a hierarchical clustering are graphically displayed in the form of a dendrogram (or binary tree), with $n - 1$ nodes.

4. In order to form discrete groups, the branches of this tree are cut at a level where there is a lot of “space”, that is where there is a relatively large jump in levels of two consecutive nodes.

5. Either rows or columns of a matrix can be clustered – in each case we choose the appropriate dissimilarity measure. It is more intuitive to show the results of clustering of variables in terms of their similarity measure, for example their correlations.

List of Exhibits