

4.3.3.1 Hierarchical clustering schemes (HCS)

An hierarchical clustering scheme takes a matrix of dissimilarity measures between a set of objects and represents the objects as being gathered into clusters on the basis of this information. It describes not one clustering but rather (for p points) p different clusterings, referred to as *levels* of a single total hierarchical scheme. At the highest level, all the objects are contained in one cluster, at the next highest there are two and so on until, at the lowest level, there are as many clusters as there are points. The defining characteristic of a hierarchical scheme is that once a point is incorporated into a cluster at a lower level it may not 'leave' that cluster at a higher one. Thus the clusters form a hierarchical scheme in the sense that each level is a special case of the next highest. We now consider in some detail the method of hierarchical clustering.

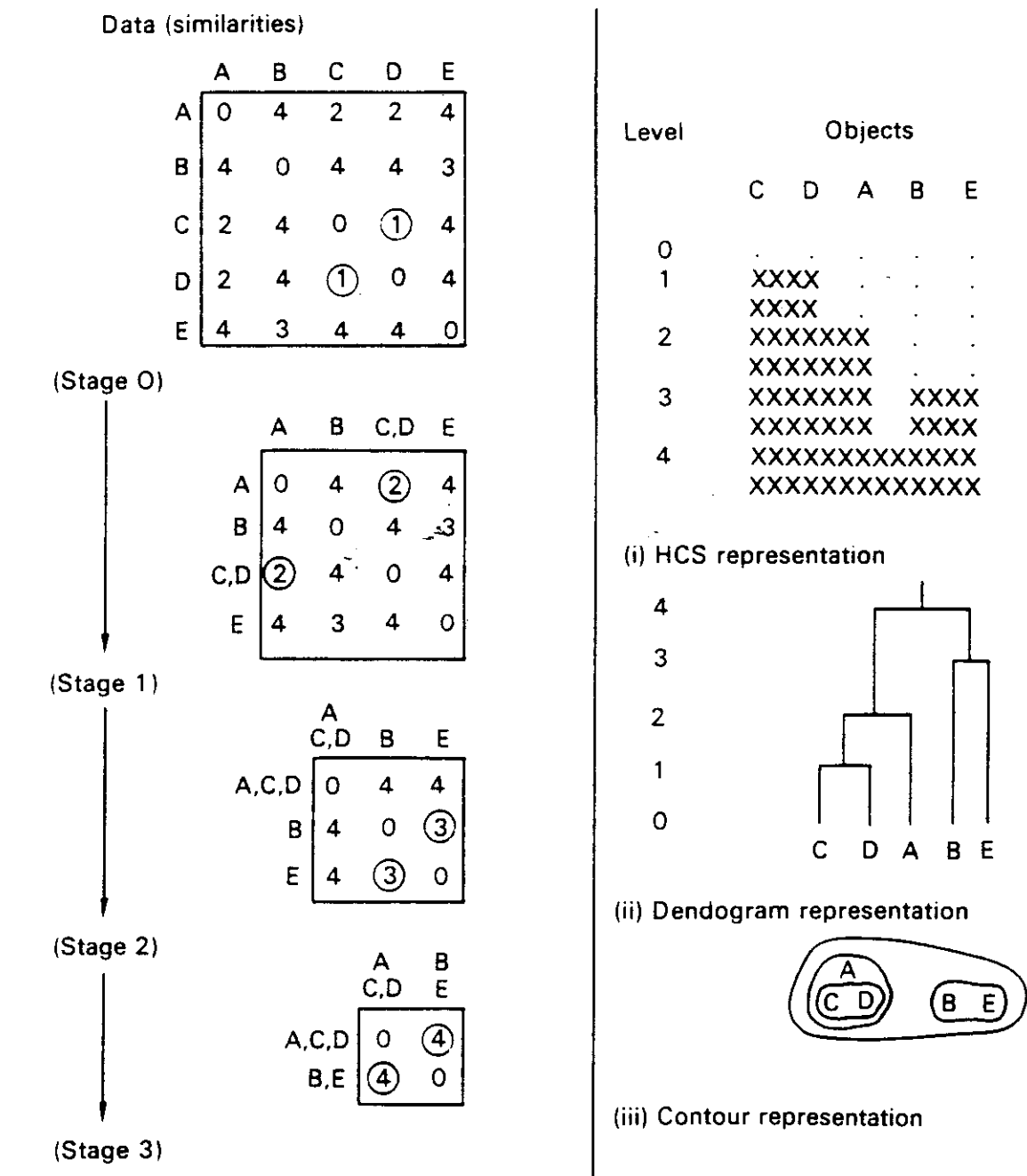


Figure 4.2 Illustrative example of the HCS procedure and forms of representation

The Method

Stage 0 The process of clustering begins by inspecting the original data matrix of dissimilarities and identifying first the most similar (or least dissimilar) pair of objects (C and D) and then merging them into a cluster. We now have the closest cluster of two points: (C, D) .

Stage 1 Points C and D are from now on treated as a single *object* and the data matrix is reduced by removing the row and column of C and D and substituting one representing the cluster (C, D) . In this example, the dissimilarity between C and each other object is the same as that between D and the same object—e.g. $\delta(C, A) = \delta(D, A) = 2$. (Normally this will not be the case.) The smallest entry in the reduced matrix, the currently most similar link, is now identified, and turns out to be between the cluster (C, D) and A . So the new structure consists of the dense cluster (A, C, D) and a set of unlinked points.

Stage 2 The new reduced matrix consists of the cluster (A, C, D) and points B and E . The smallest entry is now between B and E . This pair now form a new distinct cluster; the structure at this stage consists simply of the two clusters: (A, C, D) vs (B, E) . A new reduced matrix is formed.

Stage 3 In this final stage, the remaining two entities—the two clusters (A, C, D) and (B, E) —are merged, forming the final clustering of all the points.

The process of hierarchical clustering—forming clusters at decreasing levels of compactness—gives considerable insight into the regions of the space. In this case, we can see that the basic contrast is between the (C, D, A) and (B, E) cores of clustering.

As in other areas of data analysis, especially block modelling of social networks (see White et al. 1976 and Breiger et al. 1975) 'holes', the empty areas, frequently turn out to be quite as significant as the clusters, and both aspects have to be represented in any structural analysis. Empty regions represent two types of significant information: differentiation or dissociation between clusters on the one hand, and/or the significant absence of objects on the other hand, which might mean that certain stimuli have been neglected or overlooked in a study or that no objects actually exist which have a particular combination of attributes.

Clearly, the most significant *clustering* information is contained in the initial stages, and the most significant *dissociation* information is contained in the later stages of a clustering.

In the above example there was no ambiguity in defining the distance between a newly-formed cluster and existing objects (clusters or points), but this will not usually be the case. Consider the simplest case where we have a cluster formed of two points A and B and a third point C . There will be two distances, namely those between A and C , $\delta(A, C)$, and between B and C , $\delta(B, C)$; and we have to decide how we are going to use these to define the distance between (A, B) and C , that is $\delta((A, B), C)$. If we want the procedure to produce identical clustering schemes when the data are monotonically transformed we cannot take the obvious step of averaging $\delta(A, C)$ and $\delta(B, C)$. Johnson (1967) suggests two contrasting ways of defining this distance in this instance:

The *maximum* method (otherwise known as the diameter or complete link method) defines the distance $\delta((A, B), C)$ to be the *maximum* of $\delta(A, C)$ and $\delta(B, C)$.

The alternative *minimum* method (also known as the connectedness or single-link method) conversely defines the distance between the new cluster and the extraneous point to be the *minimum* of the distances between the extraneous point and each of the points in the cluster.

When the data satisfy the ultra-metric inequality (see 6.1.6) and are therefore perfectly representable as an HCS, the two methods produce identical hierarchical clusterings. Otherwise, the two HCSs will differ—often not markedly, but sometimes significantly.

The maximum (diameter) method picks out the largest distance within a cluster as ‘the’ distance and seeks to minimise the diameter (largest distance between the objects) within a cluster. This tends to produce a fairly small number of compact clusters.

The minimum method, by contrast, selects the smallest distance as ‘the’ distance and seeks to minimise the largest link needed to produce a chain or connected path between the objects. It tends to produce rather a large number of broken clusters and is often marked by chaining—the continued addition of a single element to a cluster.

In practice, the minimum method is usually to be preferred to the maximum method in exploring the hierarchical structure of a set of data (although both methods should be inspected to see how far the data may legitimately be represented this way.* The chief use of the HCS procedure is to examine not only relatively dense ‘local’ structure of highly proximate points (the lower levels of the clustering) but also the open or ‘global’ structure of spaces which separate or dissociate the clusters (the highest levels).

Hierarchical clustering then, possesses a number of useful characteristics:

- it presents not one, but a whole series of linked clusterings of increasing density, from a ‘clustering’ where each point is a separate cluster to the one where all the points are in a single cluster:

- it includes two commonly used types of clustering as special cases and therefore gives the user some idea of how well the data fit the assumptions of the clustering model:

- the HCS procedure is non-metric, in the sense that any ordinal rescaling of the data will produce identical results.

4.3.3.2 Clustering in high-dimensional space

The simple representation of HCS solutions within MDS configurations is only really feasible in two- or, at most, three-dimensional space. What if the user has a higher-dimensional solution and wants to gain some insight into the differential density of points in that space? An ingenious procedure is suggested by Andrews (1972; also see Everitt 1978, pp. 81–6) to represent each point as a wave form.

*Holman (1972) has shown that a set of errorless data will never perfectly satisfy both the Euclidean distance model and the hierarchical model, but will always satisfy one of the models to some extent.

6.1.6 Hierarchical clustering schemes (HICLUS)

Concisely: HICLUS (Hierarchical CLUstering) provides:

internal analysis of two-way (dis)similarity data
by means of a hierarchical clustering model,
using a monotonic transformation of the data.

Hierarchical agglomerative clustering is an unusual inclusion in a set of programs for MDS, since it really belongs in the family of discontinuous clustering ('taxonomic') models for the analysis of dissimilarity data, and the solution consists of a discrete set of groups or classification,* rather than continuous spatial models of MDS. Nonetheless, HICLUS has so frequently been used in conjunction with MDS solutions that its inclusion as a utility is virtually mandatory.

In the language of cluster analysis, HICLUS includes two agglomerative procedures (that is, methods which begin by merging two most similar points into a single grouping and continue to merge points into successively more inclusive groupings). These two methods are:

single linkage (or nearest neighbour), referred to as the 'connectedness' or minimum method in HICLUS; and

*An elementary introduction to cluster analysis is given in Everitt (1974) and a more advanced mathematical treatment is contained in Jardine and Sibson (1971). See also 8.4.2.

complete linkage (or furthest neighbour or diameter method) referred to as the 'diameter' maximum method in HICLUS.

Since the two methods define 'the' distance between a cluster and another point as the minimum or the maximum intra-cluster distance respectively, they represent two extreme ways of representing the data and neither will be changed if the data are monotonically transformed. A set of data which can be perfectly represented by an HCS will obey not only the triangle inequality requirement of any distance measure (see A2.1.1), but also the more restrictive 'ultra-metric inequality', namely:

Ultra-metric Inequality

For all triples of distinct points (i, j, k) ,

$$d(i, k) \leq \max \{d(i, j), d(j, k)\}$$

In the case of perfect data, both the 'connectedness' and the 'diameter' methods give rise to the same hierarchical clustering. The HICLUS procedure and its applications in MDS were described in detail in 4.3.3.1.

6.1.7 Maximum variance non-dimensional scaling (MVNDS)

Concisely: MVNDS (Maximum Variance Non-Dimensional Scaling) provides:

internal analysis of two-way data in a lower-triangle format of a (dis)similarity measure

by a simple distance model

using a locally monotonic and variance maximising transformation of the data guaranteed to satisfy the triangle inequality criterion.

This program and its implementation are described in detail in Cunningham and Shepard (1974). MVNDS is like all the other programs in the MDS(X) series in that it rescales the data (dis)similarities into a set of corresponding distances. It differs from them in *not* producing a configuration of points in a continuous space to which the distances correspond. Indeed, the solution distances may well not be capable of being represented spatially at all. All the MVNDS scaling procedure requires is that the solution distances have two properties. They must

- (i) satisfy the metric axioms, especially the triangle inequality; and
- (ii) be as close as possible to being a monotonic function of the data.

The basic distance model rescales the data into a set of disparities which are *as close as possible* to being distances. In any imperfect solution it is very likely that some triples of the disparities will *not* satisfy the triangle inequality, and this is one reason why they are sometimes termed pseudo-distances. In MVNDS the scaling problem is dealt with in a different way—by severing the link to a Euclidean (or other similar) space and requiring that the actual fitting values (the rescaled data) satisfy only the triangle inequality.

As they stand, these two requirements are not sufficient to obtain a (non-trivial) solution but need to be supplemented by a third, which serves the same purpose as seeking a solution in as low a dimensionality as possible in SSA: