5

Hierarchical clustering

(Seeking "natural" order in biological data)

In addition to simple partitioning of the objects, one may be more interested in visualizing or depicting the relationships among the clusters as well. This is achieved in *hierarchical* classifications in two ways: through the *exclusive* or the *inclusive* hierarchies (Mayr 1982, Panchen 1992). In the first case, clusters are arranged according to a linear ordering relation, and this sequence will be the only information added to the non-hierarchical classification. A typical example of exclusive hierarchies is represented by military ranks: a given individual may belong only to one group – indicated clearly by his shoulder strap – which is subordinate to all other individuals having higher ranks. In biology, we must go back in time to find inclusive hierarchies, the once so popular developmental sequences (*"scala naturae"*) are good examples. In the hierarchy of the animal world, the humans represented 'of course' the highest rank, followed by the apes, other mammals, marsupials, birds, and so on, the sequence concluded with the protists. In this book, we are not concerned with such arrangements any longer, and emphasis will be placed upon the other type of hierarchies. The inclusive hierarchies also involve ordering relations: small groups are nested in large clusters of objects, and these larger clusters are joined in even larger ones, and so on. Thus, an object belongs to many clusters, depending on the hierarchical *level* one is examining. In the biological sciences, such hierarchies have long been used for classification, suffice to mention the classical taxonomic ranks (species, genus, family, order, class and phylum). An inclusive hierarchy is in fact a series of partitions and can be generated by successive applications of a logical operation, the *division.* As demonstrated later, division is just one, and perhaps the least practical way of generating hierarchical classifications.

Creating inclusive hierarchies is as essential an ability of the human brain as mere partitioning. If we accept that some set of biological objects can be ordered in a natural way into a partition, then assuming hierarchical relationships for the clusters of that partition seems just as natural for us in many cases. To arrange groups into hierarchies further facilitates orientation in the surrounding world, and is by no means restricted to scientific thinking. The relatively easy interpretability and graphical attractiveness of hierarchies explain that hierarchical clustering is one of the most popular means of multivariate data exploration. In sharp contrast

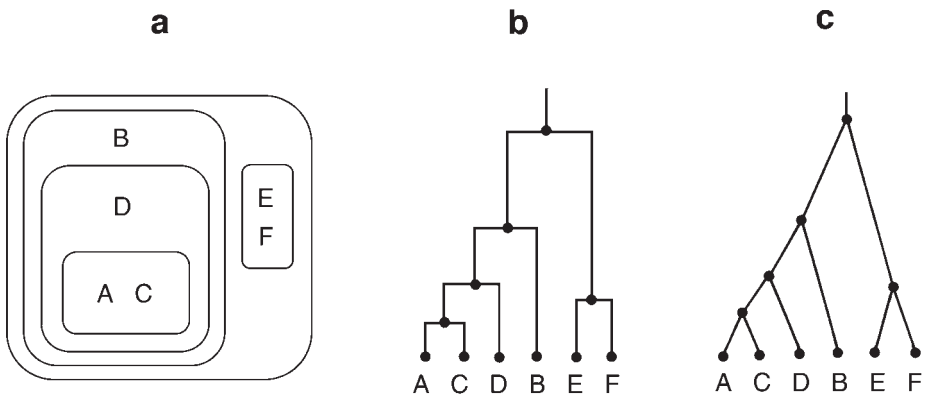

Figure 5.1. Alternative possibilities for visualizing hierarchical classifications.

with the procedures described in the previous chapter, predefinition of the number of clusters is never needed, and in fact there are only a few methods requiring specification of any parameter at all. They are therefore warmly recommended to get a quick insight into data structures – as the first step of a presumably long methodological sequence. The examples of this chapter will demonstrate that there is no universally applicable procedure that can be used in any situation, and simultaneous analyses by several methods are preferred. Yet, there is some danger that misleading results ("artefacts") are obtained by clustering (see Everitt 1980, and the examples that follow), so that their check in additional analyses by methods of dimensionality reduction (Chapter 7) is inevitable even though the final objective of our study remains to be a classification (as is in taxonomy).

An hierarchical classification can be portrayed in several ways, for example, by a nested system of contour lines (Fig. 5.1a). Such an illustration is difficult to draw and to comprehend for many objects and no measure of between-cluster relationships can be incorporated, only the topological relationships are shown. The most common[1] graphical representation is provided by *dendrograms* (Fig. 5.1b-c). A dendrogram is a tree-graph whose terminal vertices ("leaves") correspond to the objects classified[2]. Unlike contour diagrams, the dendrograms can be displayed so as to express between-cluster relationhips (distance, similarity) numerically: this is the "height" of interior vertices (hierarchical level) measured on the vertical axis. The height is best seen if the edges are broken to a right angle, as in Fig. 5.1b. The dendrogram of Fig. 5.1c is completely identical to this, although such a representation is recommended if no significance is attributed to the levels, because only the *branching pattern* of the tree is of interest (e.g., cladograms, Chapter 6). The dendrogram is a special tree graph, because it has a

1 There are further possibilities for display, such as icicle plots (Ward 1963, Johnson 1967), but these are not discussed here. The contour diagrams should not be completely forgotten, however, because in ordination spaces they may prove very useful in enhancing interpretability results (see Fig. 7.2).

2 The interior vertices cannot be identified as study objects. Such graphs are termed the *n*-trees in the literature for *n* objects (see Bobisud & Bobisud 1972), in contrast with the minimum spanning tree (Subsection 5.4.3), in which the number of vertices is the same as the number of objects. The additive trees discussed later are also *n*-trees.

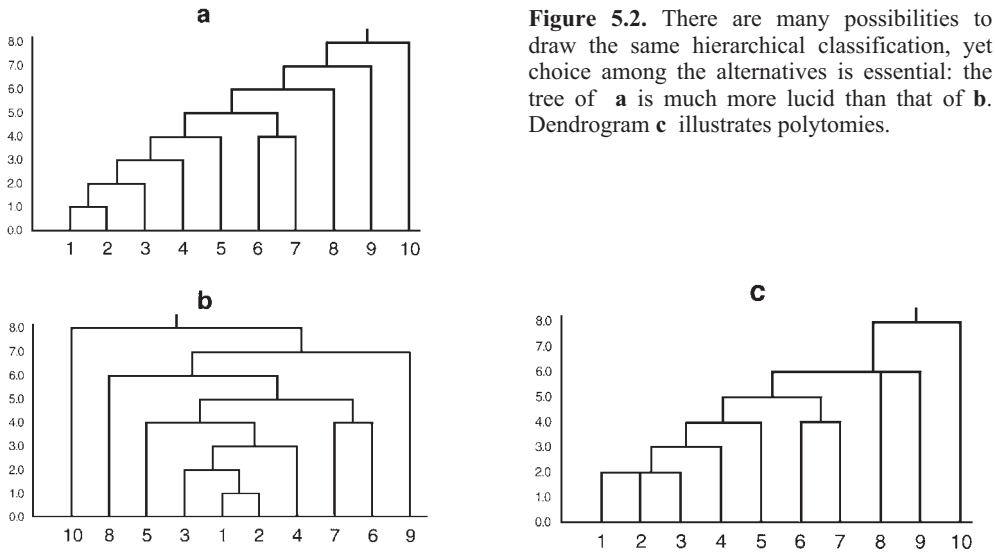

Figure 5.2. There are many possibilities to draw the same hierarchical classification, yet choice among the alternatives is essential: the tree of **a** is much more lucid than that of **b**. Dendrogram **c** illustrates polytomies.

root, the edge belonging to the vertex most distant from the leaves (as we shall see later, in Subsection 5.4.3, unrooted trees also play some role in data exploration). The dendrograms are usually drawn upside down: the leaves are in the bottom and the root is on the top, and the present book will also follow this convention. There are no strict rules, of course, because the leaves may be just as well directed upwards or the dendrogram can be positioned horizontally[3]. To some extent, the order of objects is also arbitrary; the subtrees belonging to interior vertices may be rotated, so there are exactly 2^{m-1} possibilities to display the same hierarchy. Of these alternatives, the most clear-cut arrangement is preferred, as in Fig. 5.2a, which is automatically output by most dendrogram plotting routines.

A dendrogram is *dichotomous* (bifurcating or binary) if there are three edges associated with every interior vertex (as in Figures 5.1b-c and 5.2a-b). If there is a vertex with more edges, the dendrogram is *polytomous* (multifurcating, Fig. 5.2c). Data structure and the clustering algorithm itself determine whether the resulting dendrogram is bi- or multifurcating (some cladistic methods to be discussed in Chapter 6 strive for binary trees exclusively). Appearance of polytomies within a dendrogram is indicative of special features of the data structure, e.g., equidistance of two objects from a third one.

In discussing dendrogram properties, attention needs to be paid to a new type of metrics. Any dendrogram can be written in form of a symmetric matrix **E**, in which e_{jk} is the *lowest hierarchical level* at which objects j and k belong to the same cluster. In a regular dendrogram, the following inequality satisfies for any three objects indexed arbitrarily by h, j and k:

$$e_{jk} \leq \max \{ e_{hj}, e_{hk} \} \qquad (5.1)$$

and function e implied by the dendrogram is said to be *ultrametric* (Johnson 1967). The above relationship means that for any triple of objects two of the three levels will be identical and

3 Sneath & Sokal's (1973) classical monograph provides a well-balanced mixture of these three alternative displays.

the third cannot be larger than the previous two, which is easily confirmed for the reader by looking at the dendrograms of Fig. 5.2. Condition 5.1 is obviously more stringent than the axiom of triangle inequality and is manifested in the dendrogram as a monotonous increase of hierarchical levels. Some hierarchical methods (centroid, median, see below) may violate the ultrametric inequality for some triples, and *reversals* appear in the resulting dendrogram (Fig. 5.9). This is not to say that these methods are necessarily "bad" and useless, because the centroid and median procedures do have meaningful geometric interpretation. Of course, many reversals can make the tree confusing and chaotic. Further, comparison of dendrograms is impossible with some methods if reversals are present in the dendrogram (cf. Chapter 9).

5.1 Algorithmic types

The arsenal of hierarchical clustering is extremely rich. Choice among the methods is facilitated by an – actually hierarchical – classification based on their main algorithmic features.

Agglomerative versus divisive algorithms

The process of hierarchical clustering can follow two basic strategies. The *agglomerative* algorithms consider each object as a separate cluster at the outset, and these clusters are fused into larger and larger clusters during the analysis, based on between-cluster or other (e.g., homogeneity) measures. In the last step, all objects are amalgamated into a single, trivial cluster. The strategy of *divisive* methods proceeds in the opposite way: the clustering process starts with all objects in a single cluster, which is divided in the first step into two parts. Each of them is further subdivided in the next step, and subdivisions can be continued in similar manner until every cluster has a single object (although divisions may be arrested earlier applying a stopping rule). Neither the agglomerative nor the divisive methods allow corrections: if two objects are clustered together or separated, respectively, at the beginning of the analysis, their mutual relationships cannot be changed even though at a different hierarchical level relocation would improve the classification. The classificatory ability of the human brain seems to be closer to the divisive approaches, whereas computerized realizations are much simpler for the agglomerative strategies.

Monothetic versus polythetic classifications

In *monothetic* clustering, each step of the analysis is based on a single variable, so that the resulting clusters will be identical with respect to that variable. In *polythetic* methods, decisions are always influenced simultaneously by many, possibly all of the variables involved. Therefore, the objects of the same cluster need not agree completely for a particular variable, their similarities or distances measured in the multidimensional space provide the basis for clustering. (In fact, all methods discussed in the previous chapter were polythetic). The rigorous principle of monothetic partitoning implied in earlier biological classifications (e.g., Linneaean classification of the plant world) has been relaxed considerably by the polythetic methods. Practically all agglomerative methods belong to the polythetic family; monothetic versions appear less meaningful in this regard. The divisive methods can be both monothetic and polythetic, however.

5.2 Agglomerative methods

Agglomerative clustering may follow two strategies. The *route-optimizing methods* (Williams 1971) or *d*-SAHN procedures (Podani 1989b), in which "SAHN" stands for "sequential, agglomerative, hierarchical and nonoverlapping" (cf. Sneath & Sokal, 1973), measure inter-object and inter-cluster distances (or similarities) during the analysis. That is, in each step distances are minimized or similarities are maximized. The most crucial element of these methods is the way distances between clusters are calculated (Table 5.1). Figure 5.5 demonstrates that they have simple geometric interpretation. The other family of agglomerative methods, the *homogeneity-optimizing* (heterogeneity-minimizing) procedures may also start from the same distance or similarity matrices, but the concept of inter-cluster distance is dropped. Instead, the condition of amalgamating two clusters into one is that some homogeneity criterion of this new cluster be optimal in comparison with all other possible amalgamations (*h*-SAHN methods, Podani 1989b). Such a criterion is the variance, sum of squares, entropy or within-cluster average similarity (recall Section 3.7) of the clusters (i.e., some statistic), whereas the methods cannot be interpreted in geometric terms.

Before we proceed with discussing members of these two families of methods, some of their properties need to be introduced. These details relate to the algorithmic realization of clustering, rather than to the theoretical principles. The first aspect is the space that need to be reserved in computer memory to complete the analysis (Fig. 5.3). The most economical algorithms do not require access to the original data after the distance matrix has been computed. The dendrogram is constructed on the basis of the information contained in the distance matrix only (*stored matrix approach*, Anderberg 1973; Fig. 5.3a). These algorithms are better known under the term *combinatorial* methods in the literature (Williams 1971, Lance & Wil-

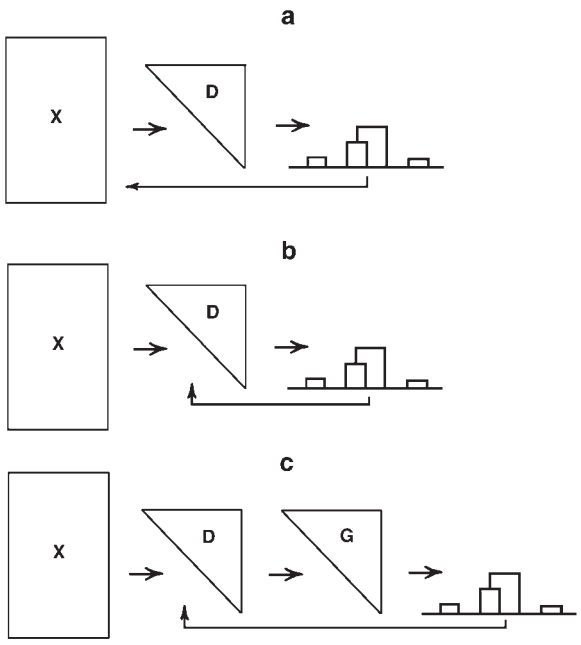

Figure 5.3. The space complexity of agglomerative clustering procedures. Matrix **X** represents the data, matrix **D** is the primary distance matrix, and **G** is a secondary symmetric matrix containing the clustering criteria computed according to **D**.

liams 1966). We have to admit that this is a somewhat misleading name, expressing that hierarchical levels pertaining to each amalgamation step are "combined" (for example, averaged or otherwise calculated) from the initial values of the distance matrix using appropriate "combinatorial" recurrence formulae. This is possible because old distances are no longer needed and can be rewritten during computations. The next group of algorithms requires simultaneous storage of the data and the distance matrix **D** (Fig. 5.3b). In each classification step, **D** is recalculated by reference to the original data (*stored data approach*, Anderberg 1973). The centroid method, for example, is known to have both combinatorial and stored data versions as well. The third group stores two symmetric matrices in memory (Podani 1989a, 1994; Fig. 5.3c), and could be called the *double matrix approach*. Having computed the distance matrix the raw data may be discarded, as in the first case. However, these distances are not used directly as clustering criteria: for this purpose a new secondary symmetric matrix is computed in each step of the analysis. An example is the minimization of the ratio of between-cluster and within-cluster average dissimilarities (Subsection 5.2.4): the second symmetric matrix contains these ratios for all the possible pairs of clusters.

The number of fusions (mergers) performed in each clustering cycle also merits our attention. One might say that upon each pass through the values of **D** the absolutely nearest pair of objects, and later clusters, are to be identified and fused into one cluster (*closest pair* or *CP*-algorithm). Some of the methods can be significantly accelerated if more fusions are allowed in every cycle: the mutually nearest pairs may be amalgamated even though their distances are far from being the optimal in the distance matrix (that is, cluster *A* is the closest to cluster *B* and vice versa; *reciprocal nearest neighbours* or *RNN*-algorithm). Bruynooghe (1978) and Gordon (1987) have shown that the *CP* and *RNN* algorithms produce identical results for several combinatorial methods (see last columns of Tables 5.1 and 5.2), and therefore use of the latter algorithm may reduce computation time considerably.

A critical and often neglected problem of agglomerative clustering is the presence of *ties* and their resolution. A tie appears if the minimum distance pertains to several pairs of objects or clusters in the same clustering pass. The chance of such events is much higher for presence/absence data than for "quantitative" variables, for obvious reasons. Many cluster analysis programs make an arbitrary choice among these pairs, sometimes srongly affecting the final result (Podani 1980, gives an actual ecological example for the presence/absence case). If one wishes to remove arbitrariness from the analysis, then the following points are worthy of attention.

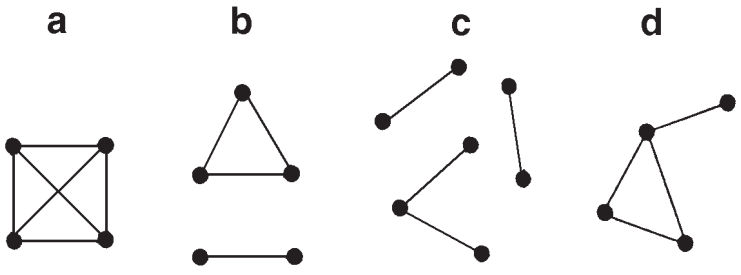

Figure 5.4. Different types of ties potentially appearing in agglomerative cluster analysis (Podani 1989a).

Ties can be best illustrated by graphs (Podani 1989a). Let us consider the objects or clusters involved in a tie as vertices of a *G* "tie-graph". Two vertices are connected by an edge if the distance between the corresponding two objects (or clusters) is the minimum in the matrix. Then, we may distinguish among four typical cases (see also Fig. 5.4):

a) *G* is a full graph, that is all vertices are connected with one another;
b) *G* falls apart into isolated subgraphs, each of them full by itself;
c) *G* is composed of isolated subgraphs, and at least one of them is not full; and
d) *G* is neither full nor it falls apart into isolated subgraphs.

In cases *a*-*b,* the resolution of ties (removal of any arbitrariness from the analysis) is quite straightforward: a *multiple fusion* amalgamates all tied objects into one cluster (case *a*) or several clusters are formed by simultaneous fusions, each cluster representing an isolated subgraph (case *b*). In the other two, more problematic cases one may follow either of the following:

– the *single linkage* resolution creates as many groups as the number of subgraphs (three and one such clusters in Fig. 5.4*c*-*d*, respectively).

– the *suboptimal fusion* ignores the tied objects, and the next lowest distance is found in **D** for which no ties appear.

Whenever the investigator suspects that the analysis does not produce unique results because of ties – a fairly reasonable assumption in the presence/absence case – the computations should be performed with and without tie-resolution, concluded by a comparison of the resulting dendrograms. Resolution by the above methods is an option in the **SYN-TAX** program package (Podani 1994). The **NT-SYS** package (Rohlf 1993a) provides the opportunity to examine all the possible dendrograms that can be obtained by arbitrary choices (the number of such alternatives can be quite high). Backeljau et al. (1996) present a review of several tree-building programs and examine whether ties are detected, and if so, how they are resolved.

And now it is due time to provide a detailed account of the most important and widely used techniques of agglomerative clustering.

5.2.1 Distance optimizing methods with the combinatorial algorithm

They start form the matrix **D** of inter-object distances or dissimilarities (if we work with similarities, they need to be converted into dissimilarity according to Formula 3.4 to ensure validity of all entries in Table 5.1). In each algorithmic step, the nearest pairs of objects or clusters are found and then amalgamated. The hierarchical level pertaining to this fusion will be shown on a vertical axis drawn besides the dendrogram. After the fusion, the distances between the newly obtained clusters and all the other clusters and objects are recalculated, whereas the unnecessary rows and columns of the distance matrix are cancelled (after amalgamating two objects, a row and a column of **D** will become obsolete). The crucial part of the analysis is the way these new distances are calculated. To complete this calculation, the recurrence formula proposed by Lance - Williams (1966, 1967a) may be used:

$$d_{h,ij} = \alpha_i d_{hi} + \alpha_j d_{hj} + \beta\, d_{ij} + \gamma \mid d_{hi} - d_{hj} \mid \qquad (5.2)$$

What we are looking for, $d_{h,ij}$, is the distance (or squared distance, Table 5.1) between the cluster newly obtained by fusing clusters (objects) *i* and *j* and another cluster (or object) *h*. d_{hi}, d_{hj} and d_{ij} are the respective distances of cluster (or object) pairs. The α, β and γ parameters

characterize the particular clustering method; they are either constants or are determined by the number of objects that previously appeared in clusters *i, j* and *h* (Table 5.1).

Single linkage (nearest neighbour) method (Florek et al. 1951, Sneath 1957). The distance between two clusters is defined as the distance between the nearest objects of these clusters (Fig. 5.5a). The method strongly emphasizes cluster separation: elongated point clouds are recognized, but clusters connected by intermediate objects cannot be detected. The internal cohesion of clusters is absolutely immaterial, and a small initial cluster can easily attract the other objects one by one in the clustering steps, leading to the so-called *chain effect.* A most appreciated theoretical advantage of the method is its insensitiveness to ties. Another property not shared by other agglomerative methods is that changes in the the resulting dendrogram will be proportional to perturbations applied to the data (Jardine & Sibson 1971).

What we have said is illustrated by the single linkage dendrograms for the two-dimensional point patterns of Figures 4.3a-f (see Fig. 5.6). The method recognizes the distinct clusters in cases **b** and **e**, irrespective of their shapes, and the detection of the three

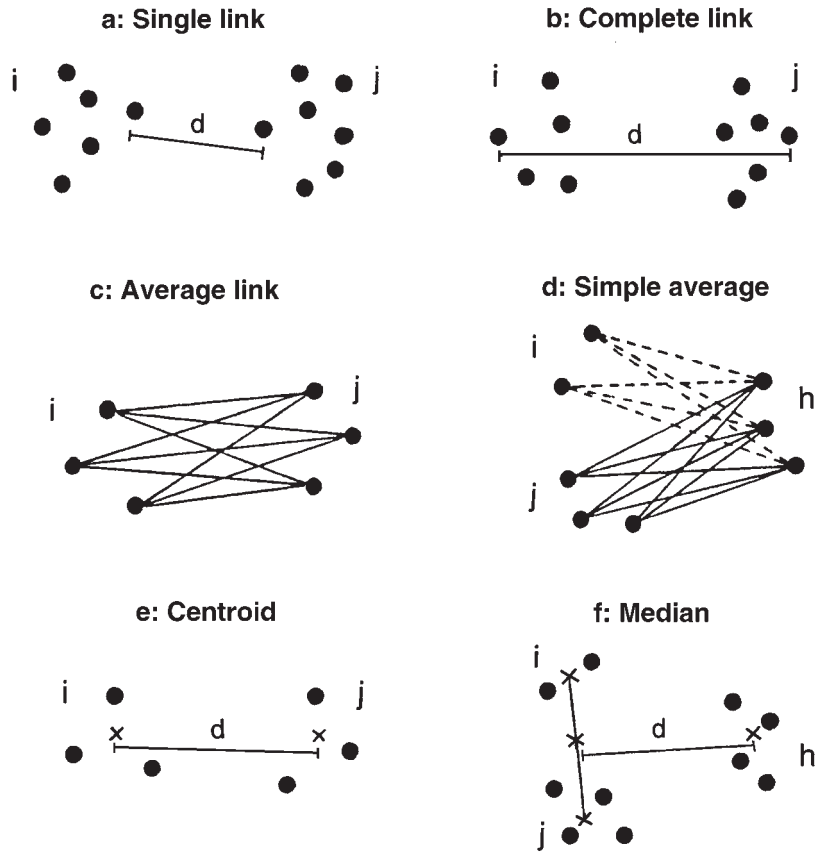

Figure 5.5. Geometric illustration of six distance-optimizing agglomerative clustering methods (after Podani 1994).

Table 5.1. Parameters and main features of distance-optimizing combinatorial methods. n_i and n_j are the numbers of objects present in clusters i and j just amalgamated.

Name	α_i	α_j	β	γ	Initial value in **D**	*RNN*- alg. applies (+)
Single linkage	1/2	1/2	0	–1/2	d_{ij}	++
Complete linkage	1/2	1/2	0	1/2	d_{ij}	++
Group avg.	$n_i / (n_i+n_j)$	$n_j / (n_i+n_j)$	0	0	d_{ij}	++
Simple avg.	1/2	1/2	0	0	d_{ij}	++
Centroid	$n_i / (n_i+ n_j)$	$n_j / (n_i+n_j)$	$-n_i n_j / (n_i+n_j)^2$	0	d^2_{ij}	–
Median	1/2	1/2	–1/4	0	d^2_{ij}	–
β-flexible	1/2 (1-x)	1/2 (1–x)	x (<1)	0	d_{ij}	–
(β,γ)-flexible	1/2 (1-x)	1/2 (1–x)	no limit	no limit	d_{ij}	–
Flexible group average	$(1–x) (n_i / (n_i+n_j))$	$(1–x) (n_j / (n_i+n_j))$	x (<1)	0	d_{ij}	–

elongated point clouds of case **d** is also almost successful (object 8 is the "problem", because it is far apart from all other points, and is therefore interpreted by the method as an *outlier*). Some "clusters" in the random case were also disclosed (**a**), but the regularity of object arrangement (**f**) is reflected in the dendrogram by the almost identical hierarchical levels. The single linkage method fails to detect the two main clusters in case **c**: the chained dendrogram reveals no more than minor, uninteresting structural details.

Complete linkage (farthest neighbor) method (Sorensen 1948, Lance & Williams 1967a). This strategy is the opposite to the previous one in every aspect; the distance of two clusters is defined as the distance between their farthest objects (Fig. 5.5b), which may be considered as the diameter (maximum within-cluster distance) of the new cluster. In this case formation of new clusters similar in size is favoured throughout the analysis, an effect opposite to chaining. As a result, the dendrograms will have a balanced shape even though the structure of data does not support this. Cluster cohesion has priority when building the tree, whereas separation of clusters is not influential.

The above characterization is supported by Figure 5.7. The dendrograms obtained for the random (**a**) and the almost regular (**f**) arrangements suggest the existence of distinguishable clusters. It may be partly right for the random case: in spite of randomness there is some tendency to form minor clusters of points. On the other hand, the hierarchy obtained for the regularly spaced objects is a typical artefact. One may confirm easily that increases of the hierarchical levels in these dendrograms are smaller than in dendrogram **b,** in which abrupt jumps indicate the separation of the four spherical clusters. However, this does not become obvious enough by looking at a single dendrogram; without a thoroughful comparison the shape of dendrograms may be completely misleading. Complete linkage analysis was able to detect the two clusters in case **c**, and the critical object 14 occurs together with the right group. The non-spherical point clouds of cases **d** and **e** were not revealed unambiguously. At three-cluster level, the dendrogram of Fig. 5.7d recalls the partition obtained by *k*-means clustering (Fig. 4.3d): the group of objects 19-25 is distinct, but the other two clusters are mixed. The group {19-25} and objects 1-7 appear together in Fig. 4.3e, showing that elongated clusters can be split easily in complete linkage analysis. To sum up, this strategy yielded acceptable results only for two of the six examples.

Group average (average linkage) method (UPGMA = unweighted pair group method using arithmetic averages, Sokal & Michener 1958, Rohlf 1963). The method is intermediate between the single and complete linkage strategies, thus attempting to compensate deficiencies of one strategy by the advantages of the other. The distance of two clusters is understood as the *arithmetic average* of all between-cluster distance values (line segments in Fig. 5.5c). During clustering, recalculation of distances must consider the number of objects previously merged in each cluster (see Table 5.1), in sharp contrast with single and complete linkage analyses which are not influenced numerically by cluster size. This will be obvious from the following example, illustrating calculations of group average clustering of five objects. The example will hopefully promote understanding the reasons why the distances and the recursion formula are sufficient for the calculations, without the original data. Let the starting semimatrix of distances, with rows and columns numbered, be given by:

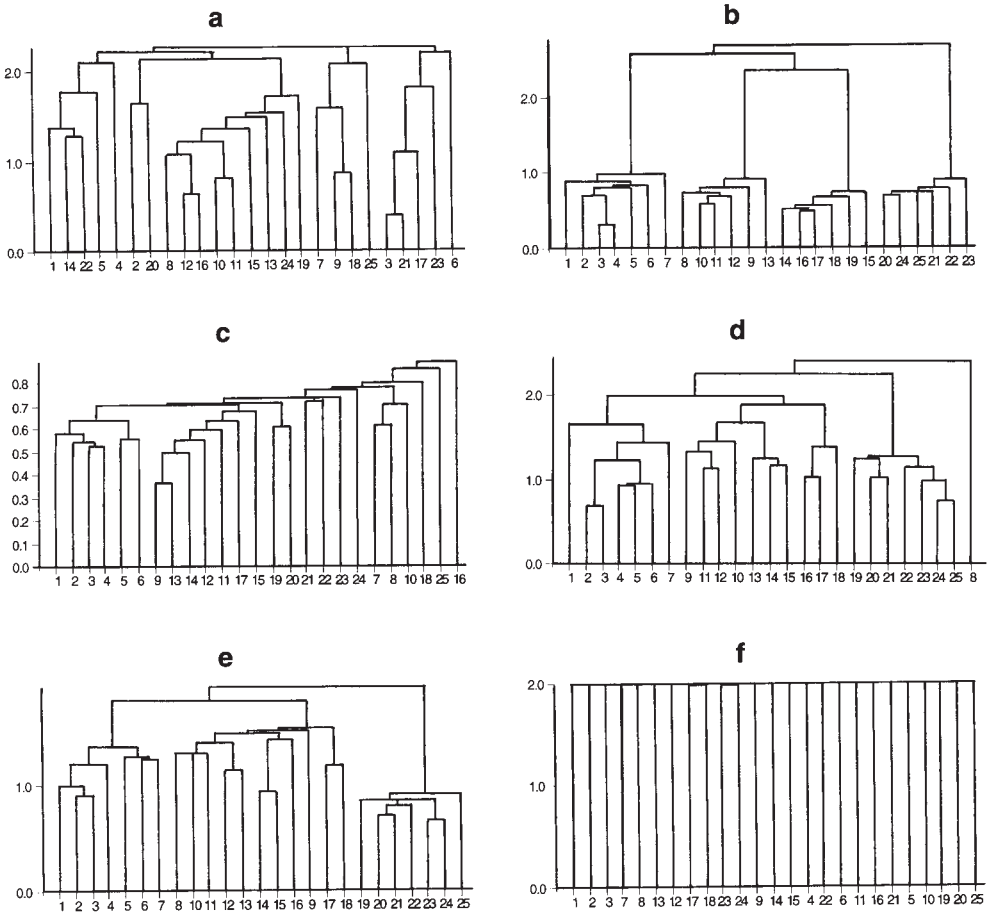

Figure 5.6. The point patterns of Fig. 4.3 evaluated by single linkage clustering. The starting matrix contains Euclidean distances among points.

	1	2	3	4	5
1	0.000	0.632	0.683	0.730	0.775
2		0.000	0.856	0.894	1.000
3			0.000	0.440	0.516
4				0.000	0.447
5					0.000

For the sake of this illustration only one fusion is performed in each clustering pass. The smallest value of the matrix is $d_{34} = 0.440$, so objects 3 and 4 are joined first. The distances between this new cluster, {3,4} with objects 1, 2 and 5 will be obtained by averaging their distances with objects 3 and 4 (see the Lance-Williams formula):

$$d_{1\{3,4\}} = 1/2 \times 0.683 + 1/2 \times 0.730 = 0.706$$
$$d_{2\{3,4\}} = 1/2 \times 0.856 + 1/2 \times 0.894 = 0.875$$
$$d_{\{3,4\}5} = 1/2 \times 0.516 + 1/2 \times 0.447 = 0.481$$

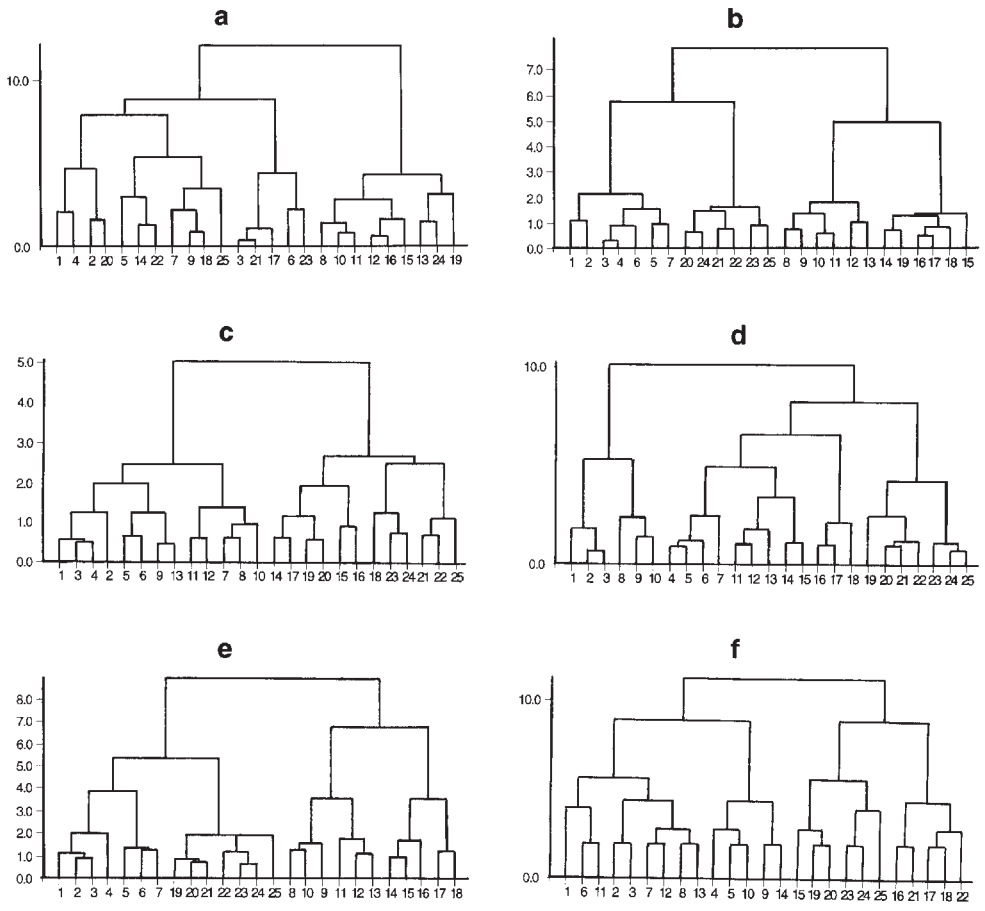

Figure 5.7. Complete linkage clustering of sample points depicted in Fig. 4.3.

The new values are written (in boldface) into the distance matrix which is reduced by one row and one column:

	1	2	{3,4}	5
1	0.000	0.632	**0.706**	0.775
2		0.000	**0.875**	1.000
{3,4}			0.000	**0.481**
5				0.000

The other values remain unchanged, of course. Upon examining the updated matrix we realize that the next smallest distance is 0.481, so that object 5 may join the cluster we obtained in the previous step at the hierarchical level of 0.481. The new cluster, {3,4,5} will take the following distances from the remaining two objects:

$$d_{1\{3,4,5\}} = 2/3 \setminus 0.706 + 1/3 \setminus 0.775 = 0.729$$
$$d_{2\{3,4,5\}} = 2/3 \setminus 0.875 + 1/3 \setminus 1.000 = 0.917$$

This is the main point here: when calculating the averages each distance value is weighted proportional to the number of objects previously fused into the clusters in question. That is, the distance of cluster {3,4} from object 1 is twice as important as the distance between objects 5 and 1. (The term "unweighted" in the name of the method is therefore somewhat misleading; it does not refer to weighting distances, but to the fact that the smaller cluster does not receive extra "weight" as compared to the larger cluster, unlike in the next method!) The new matrix is thus:

	1	2	{3,4,5}
1	0.000	0.632	**0.729**
2		0.000	**0.917**
{3,4,5}			0.000

in which d_{12} is the smallest, and therefore objects 1 and 2 will form a new cluster. (Now it becomes clear that they could have been fused earlier on the basis of their mutual closeness, because the other fusions did not modify their distance at all.) After this fusion, the matrix reduces to a single meaningful distance, obtained as:

$$d_{\{1,2\}\{3,4,5\}} = 1/2 \setminus 0.729 + 1/2 \setminus 0.917 = 0.823.$$

This new value is the topmost hierarchical level in the dendrogram. It is left to the reader to draw the result.

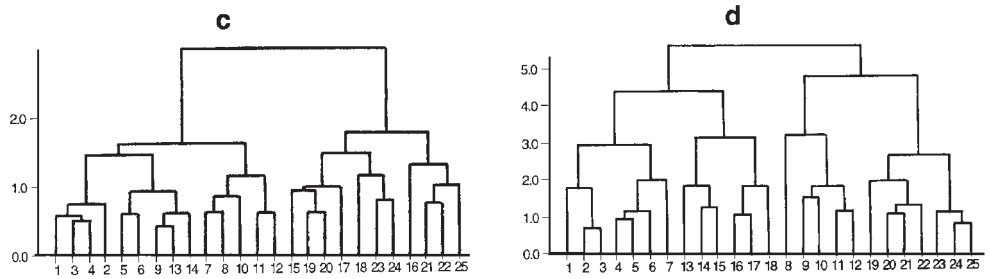

Figure 5.8. The result of group average clustering for two artificial data sets (see Figs 4.3c and d).

To save space, the outcome of group average clustering will not be displayed for all examples. In case **a**, there are only minor structural differences from the complete linkage dendrogram (5.7a), while the levels are more different, of course. For the almost trivial **b** case the four clusters are as well-separated as in Figures 5.6b and 5.7b, an expected result. The dendrogram for the adjacent two clusters of example **c** is worth showing (Fig. 5.8c), becuse it is a good illustration of the transitional behaviour of this strategy: the shape of the tree shows neither chaining nor the regular buildup of subtrees, as observed in single and complete linkage analysis, respectively. Object 14 has been attracted by the pair 9-13, unlike in the complete linkage dendrogram, so that these two results do not agree at the two-cluster level. The three elongated point clouds of case **d** (Fig. 5.8d) illustrate pretty well that a group visually judged as being intact (the middle one) can be broken into parts assigned into two other clusters. In example **e,** the inner spherical cluster is well-separated, but the surrounding arched cluster falls apart into three groups equal in size. Not surprisingly perhaps, the clustering results for case **f** are as misleading as those obtained by complete linkage analysis.

Simple average method (*WPGMA*, McQuitty 1967). This method has received much fewer applications than the previous three, because the cluster sizes are disregarded when calculating the average distances – a choice appearing quite illogical at first. As a consequence of this, smaller clusters will receive larger weight in the clustering process (hence its long name, *WPGMA* = weighted pair group method using arithmetic averages). Its geometric illustration is less straightforward, yet an attempt is seen in Figure 5.5d. Three clusters need to be considered and let us assume that clusters *i* and *j* were just merged, and its distance from a third one, cluster *h* is sought. This is derived from the pairwise distances between objects such that first we calculate the average distance between clusters *i* and *h* (for six dotted lines in the figure), and the average distance between clusters *j* and *h* (for nine solid lines in the figure). The arithmetic average of these two averages will provide the desired result, and now it becomes apparent that distances between *i-h* receive greater weight than those between *j-h.*

Sneath & Sokal (1973) point out that WPGMA clustering is a good choice when there is a reason known *a priori* to eliminate size differences between the resulting groups. For example, in a numerical taxonomic study the taxa may be represented in the sample by exceedingly different numbers of OTUs, best recognized by this weighted strategy, because UPGMA would favour clusters more similar in size.

Centroid method (UPGMC = unweighted pair group method using centroids). If the objects are assumed to be the points in the multidimensional space, then the fusion strategy of this method will be the most straightforward geometrically. Each cluster is defined in terms of the

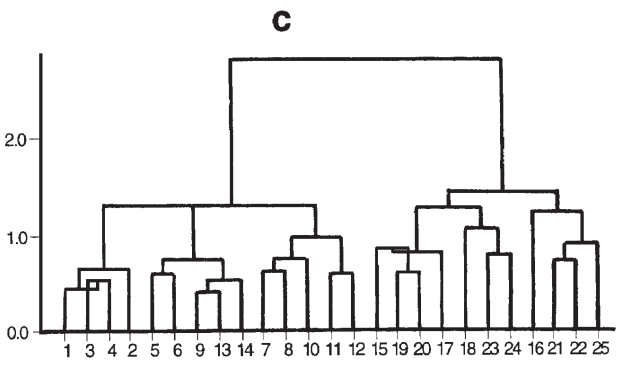

Figure 5.9. The dendrogram obtained by the centroid method often

mean (centroid) of the constituting individuals, and the resemblance between clusters is defined by the distances between the corresponding centroids (Fig. 5.5e). At a first glance, one might think that during calculations one has to retain the original data to compute the centroids, that is, the scheme of Fig. 5.3b is to be followed. Interestingly, this is not the case: Lance & Williams (1967a) have shown that there is a combinatorial version of this clustering strategy with the starting matrix containing squared distances (Table 5.1). Even though the procedure is attractive in geometric terms, there is a problem: after the fusion of clusters *A* and *B* their new centroid may get closer to the centroid of a third cluster *C* than the distance between *A* and *B*. This is manifested by the so-called *reversals* in the dendrogram, thus violating the ultrametric conditions (Fig. 5.9). Despite the potential occurrence of such disturbing events, the method serves as a good alternative to other methods whenever data averaging is meaningful for the selected coefficient and data type (as is for Euclidean distance and interval scale variables).

> The centroid clustering results for the two-dimensional sample point patterns are not reproduced here, except for case **c** (Fig. 5.9). It is left to the reader to compare this dendrogram with those obtained by the single, complete and average linkage methods.

Median method (*WPGMC* = weighted pair group method using centroids). This procedure, developed by Gower (1967) has the same relationship with UPGMC, as has simple average with group average. When computing the new centroid after a fusion, the number of objects present in the two clusters in question is disregarded, as illustrated in Fig. 5.5f. Suppose that clusters *i* and *j* with 2 and 4 objects, respectively, were amalgamated. Then, the new centroid will be obtained as the simple average of the two old centroids. As a consequence, the new centroid (preferably called the "pseudo"-centroid) will fall closer to the smaller cluster than when cluster sizes are not disregarded. As seen, the method attributes greater weight to small clusters, hence its name. Its use is recommended in those situations when WPGMA is also suitable, with the restriction that the operation of averaging data should be meaningful.

Flexible strategies. Lance & Williams (1967a) proposed a family of agglomerative methods which always produce reversal-free dendrograms provided that the following conditions are satisfied:

$$\alpha_i + \alpha_j + \beta = 1;\ \alpha_i = \alpha_j;\ \beta < 1;\ \text{and}\ \gamma = 0. \qquad (5.3)$$

For values of β close to 1, the dendrograms will exhibit strong chaining effect, resembling the results obtained by the single linkage method, whereas for $\beta = -1$ the grouping tendency is very strong, as observed for complete linkage analyses (Fig. 5.10). By changing the value of β continuously from near 1 to –1, a series of classifications is generated which may reveal much more about group structure than any particular clustering strategy applied by itself. The authors (e.g., Wiliams 1976) found empirically that the value of $\beta = -0.25$ is the "optimal" in most cases. If the values of β and γ are changed without any restriction, as proposed by DuBien & Warde (1979) then we obtain an even larger set of potential results, but many of them are hard to interpret, so the proposition by DuBien & Warde is only of theoretical interest. There is another, quite recently suggested flexible method (Belbin et al. 1992) which appears more promising. This is a variant of the average linkage method (last row in Table 5.1), and may be called the "flexible UPGMA". Based on several sets of simulated data, the authors

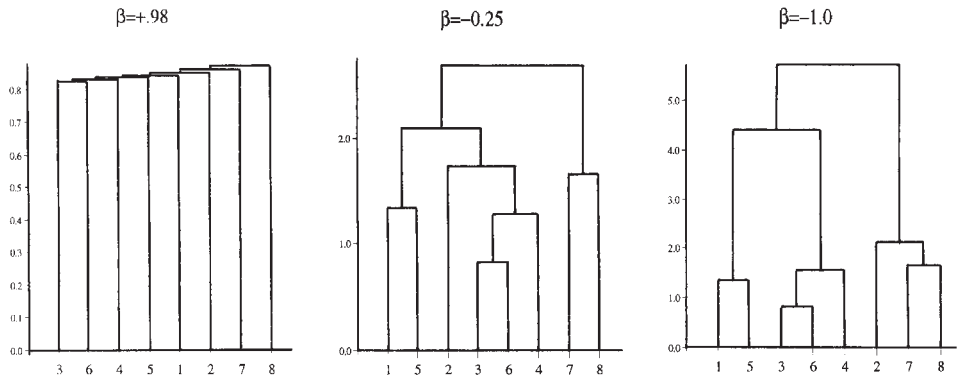

Figure 5.10. A classification series using the β-flexible method for the objects of Table A1, after standardizing variables by range. Observe the changes caused by the decrease of the β parameter. The reader may wish to evaluate the results in comparison with the Chernoff faces for the same objects as shown in Fig. 2.2.

were able to detect a narrow range of β, from –0.1 to 0.0, which reproduces inherent group structures most faithfully.

5.2.2 Homogeneity-optimizing combinatorial methods

Methods belonging to this category place emphasis upon the internal structure of clusters composed of two or more objects, by maximizing their *homogeneity* (or, which is equivalent, by minimizing their heterogeneity). The collective term homogeneity is used in lieu of a better notion, to refer to measuring overall resemblance of objects in a cluster. The sum of squared deviations, the variance and some other functions discussed already (Equations 3.105-3.115) may serve this purpose. The clustering criterion involves either the optimization of the homogeneity of the new clusters, or the minimization of the change of the homogeneity upon the fusions. These two alternatives may produce drastically different results for the same homogeneity measure; the first one usually yields clusters similar in size and is therefore less practical (cf. Anderberg 1973). Methods utilizing the variance, sum of squares and average within-cluster distance have combinatorial solutions. For these procedures, the starting **Y** matrix contains the heterogeneity measures for all the possible pairs of objects (i.e., clusters with two elements), denoted by y_{ij} for objects i and j. If i and j are fused in a clustering pass, then the recurrence equation (Jambu 1978, Podani 1979a) is used to calculate the heterogeneity that would result upon the fusion of this new cluster with any third cluster h:

$$y_{h,ij} = \alpha_i y_{hi} + \alpha_j y_{hj} + \beta y_{ij} + \lambda_i y_i + \lambda_j y_j + \lambda_h y_h \qquad (5.4)$$

(for parameters, see Table 5.2). The matrix updated this way is checked for the optimum value in the next pass. A striking difference from Equation 5.3 is that in this case each cluster, i, j and h, has its own heterogeneity measure, which is obviously zero for single objects.

Table 5.2. Parameters and main features of some homogeneity-optimizing combinatorial methods. n_i = number of objects in cluster i, $n.= n_i+n_j+n_h$, αj, λ_h and λ_j are not shown here, because they are analogous to α_i, and λ_i, respectively. $b_{hi} = \binom{n_h + n_i}{2}$, $b_i = \binom{n_i}{2}$ etc. For parameters of other methods, see Podani (1989b).

Method	α_i	β	λ_i	Initial values of **Y**	*RNN* algorithm applies (+)
Minimization of the increase of sum of squares	$(n_h+n_i)/n.$	$-n_h/n.$	0	$d_{ij}^2 \,/\, 2$	+
Min. sum of squares in new clusters	$(n_h+n_i)/n.$	$(n_i+n_j)/n.$	$-n_i/n.$	$d_{ij}^2 \,/\, 2$	++
Minimization of the increase of variance	$((n_h+n_i)/n.)^2$	$-n_h(n_i+n_j)/n.^2$	0	$d_{ij}^2 \,/\, 4$	–
Minimum variance of new clusters	$((n_h+n_i)/n.)^2$	$((n_i+n_j)/n.)^2$	$-(n_i/n.)^2$	$d_{ij}^2 \,/\, 4$	++
Minimum average distance in new clusters	$b_{hi}/b.$	$b_{ij}/b.$	$-b_i/b.$	d_{ij}	–

Minimization of the increase of sum of squares ("incremental sum of squares", Ward 1963, Orlóci 1967, Wishart 1969). Perhaps the best-known homogeneity optimizing clustering procedure in the biological sciences. Unfortunately, the method is often labeled in the literature by misleading names, such as "minimum variance clustering", which would refer more adequately to the variance-based procedures. The condition of the fusion of two clusters is that this operation causes the minimum increase of within cluster sum of squared deviations (calculated by either 3.105 or 3.106). More formally, clusters *A* and *B* may be fused if

$$\Delta SSQ_{(A+B)} = SSQ_{(A+B)} - SSQ_A - SSQ_B \qquad (5.5)$$

is the minimum of all the possible fusions in the given clustering pass. The strategy allows simultaneous fusion of reciprocal nearest neighbours, thus reducing computing time.

Since the method is both popular and easy to understand, its results are shown for all the six artificial examples. In this way the differences between distance- and homogeneity-optimizing methods become even more apparent. Comparisons are possible because in all cases – and here, too – the measurement of Euclidean distances between points is the first step. Upon examining the dendrograms of Fig. 5.11, we immediately realize that practically no importance should be attributed to the abrupt increases of hierarchical levels. In the random case (**a**) and for the well-separated four clusters (**b**), the pattern of change in levels is very similar, which is a basic feature of the method, rather than any indication of group structure and separation. (In fact, the sum of squares increases rapidly as the number of objects included in clusters increases.) Of course, there is some difference in the levels, but they would become evident from a thorough comparison of results only. In example **c**, the two main clusters are recognized such that the separation appears between objects 13 and 14. For case **d,** the method produced the "best" result we have seen thus far, because the three elongated clouds are almost perfectly recognized. On the other hand, in example **e**, the method did not outperform the previous procedures. Finally, the dendrogram obtained for the regularly spaced points (**f**) is a good illustration that greatly increased levels do not indicate by themselves any group structure in the data, and may result from cluster-free configurations as well.

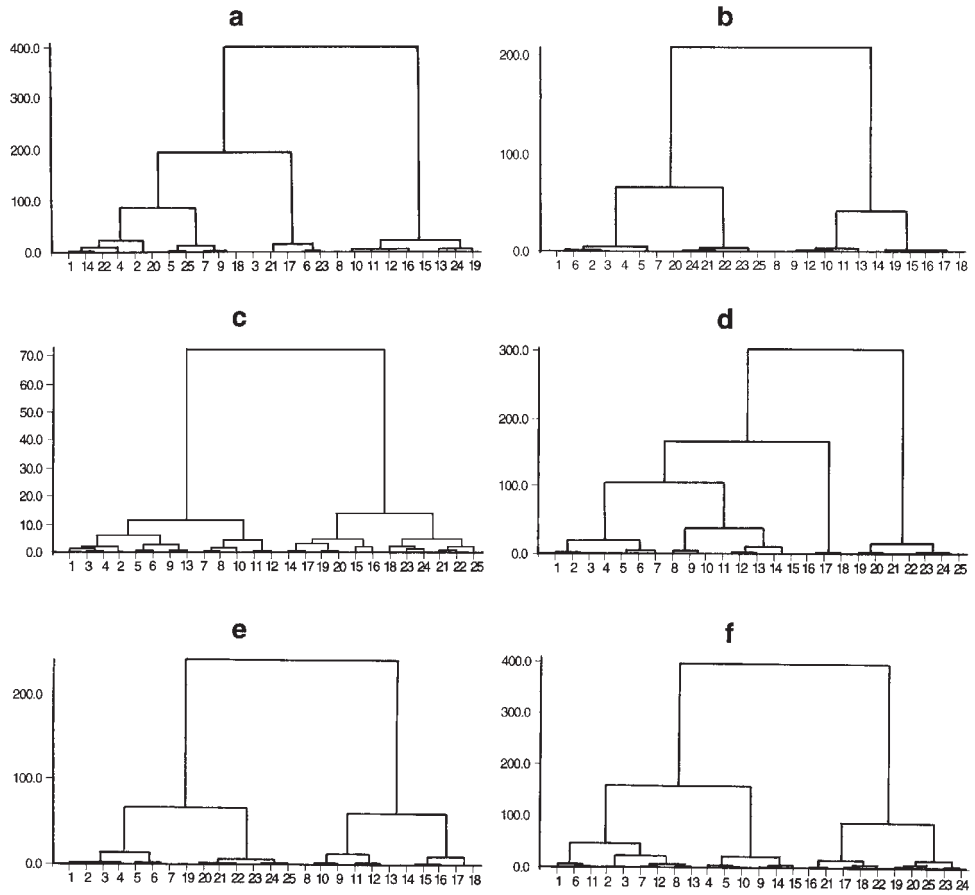

Figure 5.11. Hierarchical clustering results for the point patterns depicted in Fig. 4.3 by minimizing the increase of within-cluster sum of squares.

The y values shown for the dendrograms in Fig. 5.11 may appear contradictory with the numerical clustering results: even though increases of sum of squares were minimized, the sum of squares of new clusters are used as hierarchical levels. This was done on purpose, otherwise the diagram would be quite confusing and results obtained by other h-SAHN methods would be incomparable. In general, all h-SAHN dendrograms are illustrated using the homogeneity measures for the new clusters, no matter which clustering submodel was used.

The three clustering criteria (sum of squares, variance, mean within-cluster dissimilarity) and the fusion submodel (optimizing homogeneity of new clusters, optimizing changes) may be combined to derive further procedures for this group. Of these, methods relying upon the third criterion deserve particular attention because of its generality: any symmetric measure of resemblance can be used[4]. Sum of squares and variance are restricted to fewer cases, because of the conditions of applicability mentioned earlier (e.g., the points must be in a Euclidean

4 Originally Anderberg (1973) has discussed this method as a representative of the stored data approach, and it turned out later (Podani 1979a) that there is a combinatorial solution.

space). There is a flexible method as well, whose parameter λ may be changed within a broad range to generate classification series. For $\lambda = 0$, there is a strong chaining effect, so typical of single link dendrograms, and for large negative values of λ the resulting clusters become more and more balanced in size. The parameters of the recurrence relation and the initializing values of the matrix are presented in Table 5.2 (Podani 1989b gives a fuller account). Although the theoretical foundations are clear, the performance of these methods under controlled circumstances has not yet been examined thoroughfully, so their application on their own is not recommended.

5.2.3 Homogeneity-optimizing non-combinatorial methods

Information theory provides further means of expressing within-cluster homogeneity, but the recurrence relation is no longer valid (more precisely, no combinatorial solution has been found yet). That is, the space complexity of the method is the worst: the data must be stored during computations (Fig. 5.3a). Formulae 3.112 and 3.115 are based on presence/absence data, but their generalization to multistate nominal characters is straightforward. The best known combination of homogeneity measure and fusion submodel minimizes the *increase of weighted entropy*, that is, the condition for the fusion of clusters A and B is that the quantity given by,

$$\Delta H_{(A+B)} = H_{\{A+B\}} - H_A - H_B \qquad (5.5b)$$

be minimum in the given clustering pass ("information analysis", Williams et al. 1966). The hierarchical levels to be used in illustrating the dendrogram are the new homogeneity values. No matter which fusion submodel is chosen, the levels will increase monotonically. The algorithmic properties of these methods are not known sufficiently, however. It is unclear, for example, whether the reciprocal nearest neighbors and the closest pairs will always produce identical results (it only *seems* that they do so, but a proof is needed).

5.2.4 Global optimization clustering

All hierarchical methods introduced thus far share an important feature: for the pairwise fusion of objects (and later, clusters) they adopt a local criterion, whereas the effect of the given fusion upon the whole classification remains ignored. It is especially apparent for the distance-optimizing methods (Table 5.1), which always amalgamate the mutually closest neighbors. This closeness, i.e., the local optimum does not necessarly support a solution that is favourable for the whole set of clusters. In order to further clarify the problem, one must introduce a *measure of goodness of the classification* which can serve as a basis for finding the global optimum. There are several possibilities to define such a function, but we shall see only one, perhaps the simplest and easiest formula for the purpose, already known from the previous chapter. This is the ratio of the average within-cluster and average between-cluster dissimilarities (Equations 4.2-4.4). To recall its advantages in non-hierarchical clustering: 1) cluster cohesion and segregation are simultaneously considered, 2) being a unitless ratio, different classifications are directly comparable, and 3) any symmetric resemblance measure can be used. These advantages are effective in hierarchical clustering that adapts the following classification algorithm (Podani 1989a, see also the scheme in Fig. 5.3c):

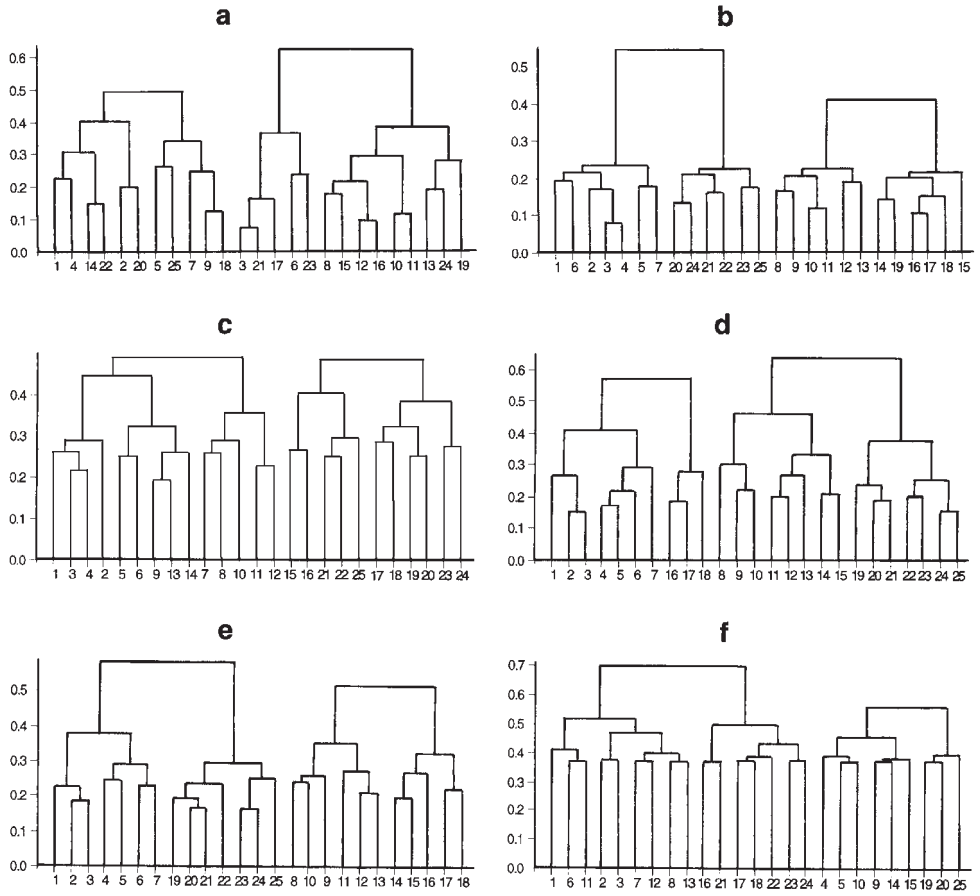

Figure 5.12. Minimizing the ratio of average within-cluster and between-cluster distances via hierarchical clustering for the point patterns of Fig. 4.3.

1) From the data matrix **X** obtain the inter-object dissimilarity matrix **D**.

2) Based on **D**, calculate the goodness of classification measure, *g*, that would result if objects (and later, clusters) *i* and *j* were amalgamated (Formula 4.4). Calculate *g* for all the possible pairs, and write them into a second symmetric matrix, denoted by **G** in Figure 5.3c.

3) The pair for which *g* is the mimimum is fused into one cluster. Only one such pair is sought, acceleration by reciprocal nearest neighbors does not work. This *g* value will be used to draw levels in the dendrogram.

4) If there are more than two clusters in a given clustering pass, the values in **G** are updated, using information stored in the original matrix **D** (see Fig. 5.3c). Then, the analysis returns to step 3. Otherwise, when there are only two clusters left, the computations stop, because for their fusion *g* cannot be calculated (there would be no between-cluster distances!). For this reason, the "dendrogram" will lack the uppermost level, so the result is in fact displayed by

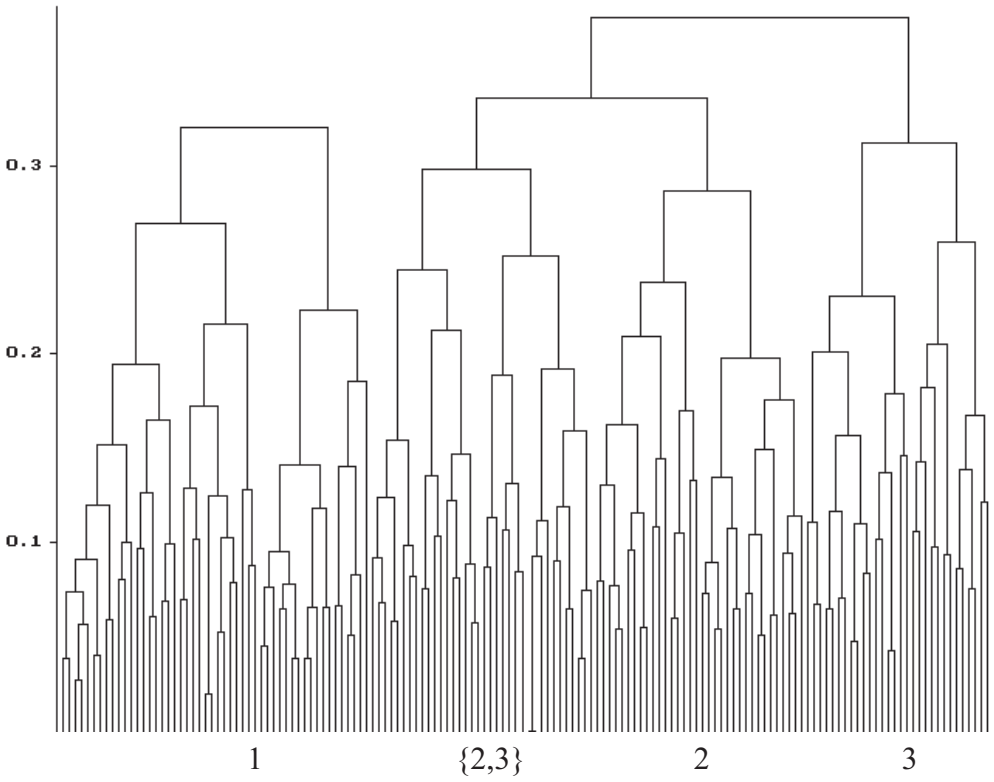

Figure 5.13. Hierarchical clustering of 150 *Iris* individuals (OTUs) using the global criterion. Numbering of objects is omitted, only major groups are identified with reference to the species (Table A2).

two subtrees, rather than a true dendrogram. Nevertheless, the classification is interpretable, because the hierarchy is complete.

To facilitate comparison with the other agglomerative strategies, the method is applied to the point configurations of Fig. 4.3. The dendrograms are displayed in Figure 5.12. In this case, levels measured on the vertical axis are worth examining, because the results are fully comparable in this regard, and not only the topological relationships are of interest. The larger the value of g, the weaker the classifiability of objects into the given number of clusters. Furthermore, within a given dendrogram differences between two subsequent levels also convey information: the larger the gap, the least meaningful the fusion in question[5]. With these considerations in mind, the six analyses may be summarized as follows. As expected, the best result yields for the four clusters in example **b**: the level is $g' \approx 0.2$, followed by an abrupt jump. The two clusters in case **c** are also separated ($g<0.5$): these are identical with those reached by non-hierarchical clustering (Fig. 4.4c). For cases **d** and **e**, the recognition of the elongated point clouds was not successful, as 'usual', and no wonder that the corresponding g value is around 0.5 for these classifications. For the random configuration (**a**), the levels increase fairly evenly, whereas for the regular pattern (**f**) the dendrogram is characterized by the narrow interval between the first and the last level, and the relatively high first fusion levels.

5 We cannot make more general statements at this point, because sophisticated procedures for the evaluation of hierarchical classifications will be discussed later, in Section 5.5.

The potential separation of the three *Iris* species is also examined by this technique. Starting from the data table (A2) the four variables were standardized by range. The clustering analysis run for minutes on a fast PC, showing the relatively high time demand of this method (analysis by distance optimizing methods would have been completed within a few seconds on the same machine). The results justify the taxonomic position of OTUs only partially: here are three clusters each consisting of OTUs from a single species, but there is a fourth cluster in which individuals of species 2 and 3 are mixed (Fig. 5.13). This is in good agreement with the fuzzy classification into three clusters, which showed that these species cannot be separated unambiguously. The relatively high overall similarity of the three species is indicated by the low topmost level (g = 0.37) and by the absence of abrupt increases. These data will be examined by other methods in the sequel, allowing comparative evaluation.

5.3 Divisive algorithms

This group comprises procedures that operate via successive bipartitioning of large clusters. Their computing demand usually exceeds the time and space complexity of agglomerative methods considerably, so divisive methods are less extensively used in practice. Nevertheless, some of them are central in importance in biological classification, being developed by statistically minded biologists in the pioneering age (early 60's) of computerized clustering and data analysis.

5.3.1 Polythetic methods

A typical and, at the same time, classical polythetic procedure was proposed by Edwards & Cavalli-Sforza (1965, see also Scott & Symons 1971) and is mentioned first as an illustration of the underlying theory. The subset A of objects is subdivided in a given step into clusters $A1$ and $A2$ if the quantity

$$\Delta SSQ_A = SSQ_A - SSQ_{A1} - SSQ_{A2} \qquad (5.6)$$

is maximum. In words: the sum of squares for the new clusters must decrease as much as possible after the division. According to the original proposition, all the possible divisions need to be examined to find the optimum. However, for practical sample sizes this would be a formidable task, because the number of possibilities to examine is astronomical, and even the fastest supercomputers cannot help (recall the number of ways m objects can be clustered into two groups, Formula 4.17 and see more details below). The classical Edwards - Cavalli-Sforza algorithm is suitable to no more than 20-30 objects. There is an approximate solution for large data sets achieved by the branch and bound algorithm[6] described by Chandon et al. (1980). Its practical implementation in commercial software packages is a timely task.

Classification based on ordinations. A group of polythetic clustering methods involves indirect classification; first an ordination of objects is obtained and the ordination scores are used subsequently to derive the groups. To understand the essentials, therefore, one has to go ahead in this book and see how ordination axes are constructed. The best known technique of this sort has the acronym TWINSPAN ("Two-Way INdicator SPecies ANalysis, Hill 1979a), which could be referred to under a different term, a possibility raised by the author himself ("dichotomized ordination analysis"), which better reflects the origin of the procedure. The TWINSPAN method is most popular among vegetation ecologists and phytosociologists. The

6 This term will reappear in the context of cladistic tree optimization in Chapter 6.

centroid of coordinates for the first correspondence analysis axis (section 7.3), which explains the highest proportion of variability among all axes, is calculated first. Objects falling to the left and to the right of the centroid will constitute the first two main clusters, which can be refined with some relocations of objects that fall close to the centroid. The clusters are then further subdivided into smaller and smaller groups. Since the objective of correspondence analysis is a simultaneous ordination of variables and objects, the variables (usually species) will also be classified in similar manner, allowing preparation of a rearranged data matrix which shows mutual correspondence of object and variable groups. This topic will be elaborated at length in Chapter 8.

The proposition that ordination axes can be used for divisive clustering predates the first description of TWINSPAN, however. Lefkovitch (1976) pointed out that a principal coordinates ordination (Subsection 7.4.1) from a matrix **E** of ultrametric measures (5.5.1) can be used to reconstruct the hierarchical classification: along each ordination axis the sign of coordinates will determine the group membership. Division along the first axis gives the first two clusters, their subdivision along the second axis yields four clusters, and so on. If it is possible, then why not to revert the whole strategy and start the analysis with the matrix **D** of inter-object distances, and examine the sign of coordinates for the resulting ordination axes to create a divisive clustering[7]? Williams (1976) proposed the fairly similar POLYDIV procedure which is based on principal components analysis such that the set of objects is subdivided into two groups to maximize decrease of within-cluster sum of squares. To be correct historically, this suggestion was first raised in the paper by Lambert et al. (1973).

5.3.2 Monothetic divisions

Monothetic classification has once been exclusively used to reveal group structure in the data. Some of its algorithmic realizations did not require the computer, although hand calculations were quite cumbersome. Biologists, especially vegetation scientists know very well the method of *association analysis* developed for clustering from presence/absence data. The most widely used version of association analysis is due to Williams & Lambert (1959, 1960) who made Goodall's (1953) original algorithm more efficient and operational. The idea is that the variable which has the highest "association" with all the others or, in other words, which is the most informative variable in the data set is identified. Based on the presence and absence of this variable, the objects are divided into two groups, and then the classification is refined by finding further divisive variables separately in these groups. The analysis may be continued "down" to the objects, although small details of such classifications are uninteresting and less reliable. The most critical part of the analysis is the choice of the mathematical function for finding the divisive variable. In the first versions of the algorithm, pairwise χ^2 scores (see Formulae 3.14-15) were calculated for all pairs of variables and then summed up for each variable. The divisive criterion was thus given by:

7 Lefkovitch was in doubt whether such a classification is polythetic or monothetic. Since the divisions are based on an abstract, synthetic variable, rather than on a single original variable, the method can be safely labeled as being polythetic.

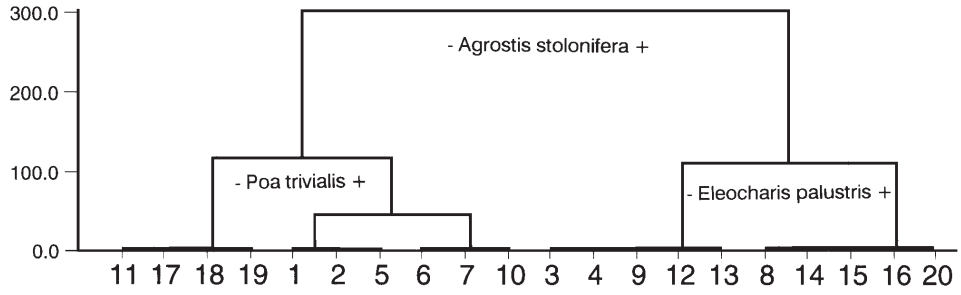

Figure 5.14. Association analysis from the dune vegetation data (Table A4) using the divisive criterion given by Equation 5.9. Only the first three divisive species are indicated. On the vertical axis, pooled entropy is measured. Compare this classification with Figure 7.17.

$$\max_i \sum_{j=1}^{n} \chi_{ij}^2, \qquad i \square j\,. \quad i \neq j \tag{5.7}$$

A disadvantage of this formulation is its inapplicability to low cell frequencies of the 2\2 contingency table, so that many variables had to be omitted from the analysis. There are two solutions of this problem, which do not lead by necessity to the same result. The χ^2 function may simply be replaced by the corresponding information theory formula (Podani 1979b), which takes the following form for the 2\2 table:

$$I = m \log m + a \log a + b \log b + c \log c + d \log d - (a+b) \log (a+b) - (a+c) \log (a+c) - (b+d) \log (b+d) - (c+d) \log (c+d) \tag{5.8}$$

("mutual information" of variables based on $m = a+b+c+d$ objects). The formula applies without restrictions as to cell frequencies, so that variable deletions are unnecessary. The relation $2I \rightarrow \chi^2$ holds true (cf. Kullback 1959, Orlóci 1978) in general, so I may be used efficiently as the divisive criterion. Otherwise the algorithm is the same as proposed by Lambert & Williams: the mutual information values are added for each variable and the one providing the maximum sum is found. The set of objects is subdivided according to this variable, and then new divisive variables are identified in the resulting clusters, and so on. The subdivisions are best stopped after a predefined T threshold is reached. On rare occasions reversals may appear in the resulting dendrogram: that is, increases of hierarchical levels are not monotonic.

The other possibility to replace χ^2 is to find the variable whose presence and absence defines two subclusters $A1$ and $A2$ such that the quantity

$$\Delta H_A = H_A - H_{A1} - H_{A2}\,, \tag{5.9}$$

that is, the decrease of *pooled entropy* ("information fall", Williams et al. 1966, Lance & Williams 1968) is the maximum. H is calculated using Formula 3.112. The levels of the dendrogram are actual within cluster entropies, rather than entropy changes.

An obvious advantage of monothetic clustering is the direct interpretability of results. The groups can be easily characterized in terms of the divisive variables, and the hierarchy implies an identification key to the groups to which new, as yet unclassified objects may also be assigned (although in the latter case we must forget that addition of new objects will change

more or less the association structure of variables). On the other hand, the rigorous monothetic principle may give rise to serious *misclassifications*: it can very well happen that an object is assigned to group *A*2, even though it has a high overall resemblance to group *A*1, except for the divisive variable. In vegetation science, for example, which has been a noted field of application of association analysis, a species may be absent from a site just by chance, even though the other species indicate that it "should be" there. Monothetic classifications are therefore often improved afterwards by a relocation algorithm, which corrects for most of such misclassifications (Crawford & Wishart 1968, Weir 1970).

> Association-analysis performs the best for large data sets with many variables, because association values among variables are statistically more reliable. Whenever possible, the number of variables should be larger than the number of objects. Although the sample data set of Table A2 does not fulfil this requirement, it is used here to illustrate the procedure. Both information theory criteria led to the same result at high levels, and it is therefore sufficient to show only one of them (Fig. 5.14).

5.4 Special clustering procedures

This section introduces alternative clustering algorithms that do not fit the above methodological framework, yet they can be used to generate hierarchical classifications. The subject matter is highly diverse, however, and the forthcoming discussion serves as an illustration of techniques of the author's own, and somewhat biased, choice only. Note, therefore, that the topic of hierarchical clustering is very far from being completely exhausted in this book.

5.4.1. Constrained clustering

The surveyor may want to impose some external criteria upon the process of classification by allowing that the clusters need not reflect conceptual inter-object distances faithfully. Instead, distances or neighbourhoods in the *field* will be decisive, as achieved by procedures of constrained classification. In palinology and paleontology, for example, only neighbouring strata may be allowed to get together during clustering so as to incorporate stratigraphic information as well. That is, the sequence of strata is not represented directly in the data, yet it influences greatly the result. The reasoning of such constraints is that two remote strata should not get together, even though their species composition is quite similar. In addition to temporal information, spatial relationships in the real two- or three-dimensional world may also be considered. All the methods discussed earlier in this chapter may serve the goal of constrained clustering, provided that the algorithm is modified appropriately.

In order to understand the necessary modifications, let us introduce an adjacency graph with vertices as the objects we wish to classify. Two vertices are linked by an edge (or arch) if the corresponding objects are connected in space or time, or are associated in some other way. When running an agglomerative algorithm, the distance matrix and this adjacency graph are inspected together. Distance values that do not have corresponding edges in the graph are simply ignored, and the fusions are based on the remaining distances. After merging two objects, they will appear as a new, single vertex in the reduced graph and will keep all their original relations. The single linkage method (Gordon & Birks 1972) and the incremental sum of squares algorithm (e.g., Grimm 1987) are the most common procedures modified this way.

Constrained classifications may be obtained by divisive clustering as well, but in this case the objects must be linearly ordered (in space or time) by the constraining principle. The objective is to remove that edge from the graph first for which the decrease of within-cluster sum of squares is the maximum. For m objects only m-1 divisions need to be examined, a very reasonable number compared to the complexity of many other classification algorithms, because this is the number of possible cuts that can be applied to the sequence. Then, the small clusters are subdivided further using similar criteria. For presence/absence data, one may also maximize decrease of the pooled weighted entropy (Formula 3.112), as suggested by Gordon & Birks (1972). The hierarchy need not be completed: after reaching the desired number of clusters the analysis may be stopped. Then, the current partition of objects may be improved by relocations using a constrained non-hierarhical clustering algorithm (Birks & Gordon 1985).

5.4.2 Adaptive clustering

The clustering procedures discussed thus far will always produce some final result, a classification, even though the method chosen is not suited to detect the inherent structural properties of the particular data set. There were a wide variety of examples to convince the reader that the different methods are sensitive to different aspects (different cluster shapes, for example) of the data, so that careless analyses may lead to false interpretations. Much benefit is gained from a simultaneous classification by many methods and subsequent evaluation of results (a possibility expanded later in this book), but there are other propositions. We may construct, for example, an algorithm that checks for the presence of some typical situations in a preliminary scan of the data and then, according to the results and possibly by some intervention of the user, will modify itself to the properties of the data. In other words, the classification procedure adapts its own strategy to the case examined. Of the several attempts to define adaptive clustering schemes the method proposed by Rohlf (1970) merits particular attention. The user prespecifies some point cloud shapes which will certainly be recognized by the program. The generalized distance (3.95), for example, favours elliptical clouds. Then, the distance between point i and a cluster j is obtained such that matrix **W** is calculated based only on objects that are already present in cluster j, whereas the distance between the same point and cluster l is computed using matrix **W** for objects of this latter cluster. In other words, the inherent structure of already existing clusters is decisive and, in this sense, the classificatory process is automatically adapted to classes already formed. Another procedure ("*mode analysis*", Wishart 1969) operates by filtering noisy elements (see Fig. 4.6) first in order to best recognize dense point clouds whose shapes are not specified in advance. The user specifies, however, an integer k and a radius r at the outset. During the analysis, each point is examined for the presence of at least k other points within distance r. Those points for which this condition holds true are evaluated by single linkage clustering, whereas the others, as noise elements, are discarded. By changing the value of r successively, one generates a series of classifications in which the number of classes first increases, and then starts to decrease (finally, for sufficiently large r we shall have a single cluster). Further details on adaptive clustering are given in Sneath & Sokal (1973: 212-214) and Gordon (1981:137-139). As Gordon notes, sooner or later we shall have the opportunity to develop an interactive algorithm which will continuously require corrections and decisions by the user, thus representing a breakaway from fully automated clustering.

5.4.3 Minimum spanning trees

Besides dendrograms, there are other types of graphs that may prove useful in revealing and illustrating group structure in the data. First of all, the "*minimum spanning tree*" needs our attention, which differs considerably from conventional dendrograms. A striking difference is that each vertex corresponds to an object, so there are no "abstract" vertices in the graph. For *m* points there are therefore *m*-1 edges, each weighted by the corresponding distance value. There are no circles, of course, and – what is perhaps the most important of all of its properties – the sum of the edge lengths, the weights, is the *minimum* (Gower & Ross 1969, Rohlf 1973). The minimum spanning tree is derived from the distance matrix of objects such that in each step a new edge is defined between the nearest two points (representing two objects) if the introduction of this edge does not give rise to a circle in the graph. That is, the smallest two distances will always have associated edges, so we may begin the search with the third smallest distance. When the number of edges reaches *m*–1, the analysis stops. The method has close relationship with single linkage classification: divisive clustering based on the graph will result the same hierarchy as the one produced by single linkage analysis. By successive removals of the longest edge from the graph, the resulting subgraphs can be identified as single link clusters (Gower & Ross 1969).

Figure 5.15 shows the minimum spanning tree fitted to the points of Figure 4.3a. The three longest edges are marked (a, b and c), so that one can easily verify that removal of these edges will produce the main clusters of the dendrogram of Fig. 5.6a. However, while the dendrogram does not show the pair of objects that are actually "responsible" for the appearance of an edge, the minimum spanning tree does. When one wishes to identify these objects, a single linkage analysis followed by generating a minimum spanning tree is the admissible strategy.

Notwithstanding its relationships with clustering, this tree is best suited to other kinds of problems. Minimum spanning trees are essential in checking the general validity of two-dimensional ordination displays (Digby & Kempton 1987:99, Gordon 1981:155, Dunn & Everitt 1982:75, and others): the tree projected onto the ordination of objects will show whether the relative closeness of object pairs is an "artefact" or, in other words, the two dimensions are sufficient to portray the distance structure of objects faithfully. If the route between two neighbouring points runs through some others falling far apart, then we have a

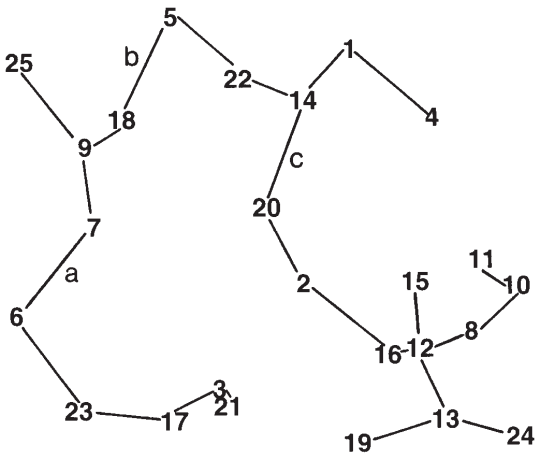

Figure 5.15. The minimum spanning tree fitted on the points of Fig. 4.3a.

good reason to assume that the two dimensions actually shown are not enough (see also Subsection 9.5.2). Rohlf (1975a) suggests that the tree may also used to detect outliers in the data set.

5.4.4 Additive trees

Dendrograms and minimum spanning trees do not exhaust all possibilities of our graph theoretic tools of representing distance structures. The rigorous ultrametric relationships implied by dendrograms do not necessarily reflect well the inter-object distances. To see this point, let us consider the following semimatrix of distances among five objects:

```
0.0 12.0 23.0 30.0 32.0
     0.0 25.0 32.0 34.0
          0.0 31.0 33.0                                    (5.10)
               0.0 20.0
                    0.0
```

Starting from this, group average clustering will produce the dendrogram of Figure 5.16a. The distances between objects 1 and 2, as well as between 4 and 5 are not distorted by this tree, but for the pairs of 1 – 3 and 2 – 3 the distance will be identical, 24.0, even though the original values in the matrix were 23.0 and 25.0, respectively. This fact did not bother us very much when classification was our objective, because we did not attribute too much importance to hierarchical levels. In addition to dendrograms, however, one may wish to generate trees that attempt to retain the original distances for all object pairs as much as possible. This goal is achieved by the so-called *additive trees* which have long been used in psychological data analysis (Sattath & Tversky 1977, Shepard 1980, Pruzansky et al. 1982). For the above matrix, the additive tree is given in Fig. 5.16b, which immediately shows the greatest difference from dendrograms: the objects are not aligned to a single horizontal line. By careful examination of the tree, one may confirm that the distance for any two objects is obtained by adding the length of edges along the route between the corresponding two vertices (this is the *patristic distance*, Farris 1967). Since levels are not drawn, the usual dendrogram-like shape may be replaced by conventional graphs, as in Figure 5.16c, which does not emphasize classifications any longer. Indeed, the next chapter will demonstrate that the primary function of additive trees is not classification.

The above example was constructed to be so elegant on purpose; actual distance matrices can rarely be represented perfectly by additive trees. For certain pairs, the distances are smaller than the sum of edges, whereas for others the distances are more or less larger. A certain type of distortion appears here, but this is not the same as the distortion implied by dendrograms. The algorithm for constructing additive trees is not as simple as the algorithm of group average clustering or other hierarchical methods, and the reader is referred to Sattath & Tversky (1977) for a full account. (The neighbor joining technique to be discussed in the next chapter gives a fairly good approximation to the results obtained by the Sattath-Tversky algorithm.) Instead of giving the computational details, it seems worthwhile to expand the discussion of two other properties of additive trees. Whereas the matrix **E** for dendrograms, as we have seen when discussing formula 5.1, satisfies the criteria for being an ultrametric, the matrix **A** of patristic distances within the additive tree meets the condition of the so-called *four point metrics*. This postulates that, irrespective of indexing the points, for any four of them the following *additive inequality*

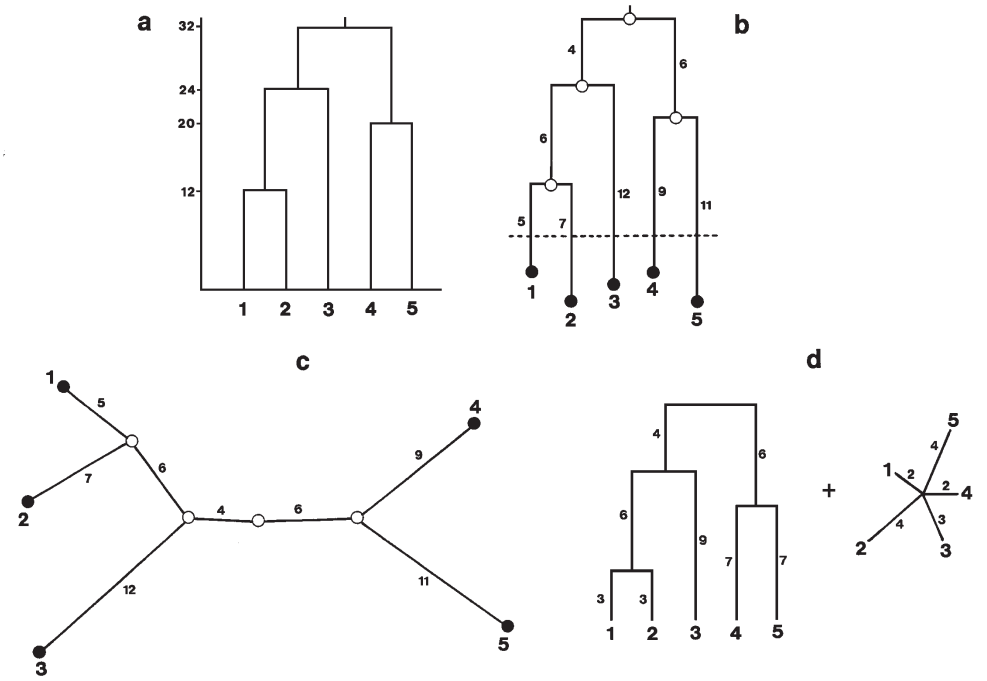

Figure 5.16. The group average dendrogram (**a**) and an additive tree (**b)** derived from matrix 5.10. The additive tree may be decomposed into a dendrogram (**c**) and a star tree (**d**).

$$a_{hi} + a_{jk} \leq \max \{ a_{hj} + a_{ik} , a_{hk} + a_{ij} \} \qquad (5.11)$$

holds (Buneman 1971, Patrinos & Hakimi 1972, Sattath & Tversky 1977). If the four points are conceived as the vertices of a tetrahedron such that their distances (a total of six) are proportional to its edge lengths, then the three sums for the opposite edges will produce an isoscele. If matrix **D** satisfies the above criterion, then it is automatically a patristic distance matrix, otherwise **D** can only be approximated by some **A** patristic matrix.

Another important property of additive trees is illustrated in Figure 5.16d. Every additive tree can be decomposed into a dendrogram and a star tree (or a "bush'): the tree cut at the level shown by a dotted line in Figure 5.16b becomes a dendrogram, whereas the pruned terminal branches will give a bush (Carroll 1976). This means that every dendrogram is an additive tree with a zero bush component, so that the ultrametric inequality is a stronger condition than the additive inequality. The different inter-point measures can thus be arranged into the following inclusion sequence (Le Calvé 1985): *dissimilarity* $\subset$ *metric* $\subset$ *Euclidean distance* $\subset$ *four-point metric* $\subset$ *ultrametric,* that is, the first is the most general, and the last one is the most special measure.

5.5 Evaluation of hierarchical classifications

The illustrative examples were supplied to convince the reader that the results of hierarchical clustering do not stand on their own in most cases, and further work is necessary to validate the classification obtained. The most optimal strategy of the data classifier is to evaluate the hierarchies obtained, supplemented by comparisons of alternative classifications of the same set of objects, and simultaneous analyses by methods of ordination. This section will discuss some possibilities of evaluating dendrogram properties. (For comparisons we need at least two results, of course, and the topic will be elaborated in much detail in Chapter 9.)

5.5.1 Measures of distortion

Evaluation of hierarchical classifications may emphasize very different aspects of the results. Firstly, let us consider *distortion*, a common feature of all dendrograms already mentioned in Section 5.4. The pairwise distances implied by the dendrogram, that is, the ultrametrics, may greatly differ from the original distances: it is extremely unlikely, yet not impossible, that original distances are at the same time ultrametrics. When the starting distances are converted to ultrametrics, the distance structure is inevitably distorted. A clustering method performs well if the change in the direction $\mathbf{D} \otimes \mathbf{E}$ is small. To measure the extent of this change, the linear correlation (Formula 3.70) has been widely used, under the name *cophenetic correlation* (Sokal & Rohlf 1962). Comparison of matrix **E** with matrix **D** from which it was derived represents a special case of matrix comparisons to be discussed in Subsection 9.2.1. The correlation for these two matrices is given by:

$$COPH_{(\mathbf{D},\mathbf{E})} = \frac{\sum_{j=1}^{m-1} \sum_{k=j+1}^{m} (d_{jk} - \bar{d})(e_{jk} - \bar{e})}{\sqrt{\sum_{j=1}^{m-1} \sum_{k=j+1}^{m} (d_{jk} - \bar{d})^2 \sum_{j=1}^{m-1} \sum_{k=j+1}^{m} (e_{jk} - \bar{e})^2}} \qquad (5.12)$$

The diagonal values of both matrices are ignored and, due to the symmetry property, the comparison may be restricted to the lower semimatrices. The cophenetic correlation is mostly favoured in numerical taxonomy, although any classification studies may benefit from such comparisons to select the method implying the lowest distortion. The value of *COPH* usually ranges between 0.6 and 0.95, and the best fit to a given matrix is achieved by group average clustering in the majority of cases (Sneath 1966, Boyce 1969, Sokal & Rohlf 1970).

Let us try cophenetic correlation to compare the single link, complete link and group average methods for the sample data sets **b** and **c** (using the respective dendrograms in Figures 5.6-8). In case **b**, when we have four obvious clusters, the best result was produced by average linkage clustering (*COPH* = 0.845), although the performance of the other two strategies is only slightly worse (for single link, *COPH* = 0.839 and for complete link, *COPH* = 0.830). For the adjacent clusters (example **c**), the sequence of methods changes to average linkage, complete and single linkage and the coefficients are much lower (correlations are 0.735, 0.729 and 0.426, respectively).The low value for the single link strategy is expected, because we know already that the method is misled by the absence of gaps between clusters. The incremental sum of squares method is neglected in the comparisons for good reasons: the hierarchical levels correspond to sum of squares, rather than distances, and their correlation with ultrametrics, though formally computable, is not commensurable with the above results. The

problem of incomparability occurs for the global optimization method as well, further complicated by the absence of the topmost hierarchical level, so that any attempt to calculate the correlation would fail. One must always check for possible incompatibility and incomparability of dendrograms and matrices to avoid false conclusions drawn from misuse of the cophenetic correlation function.

In addition to *COPH*, other formulae can also be used successfully, such as squared Euclidean distance (Hartigan 1967) or Kruskal's (1964) stress function. The latter one will be introduced in the discussion of multidimensional scaling, the main field of application of the function (Subsection 7.4.2). The problems with non-comparable levels can be alleviated using rank correlations (Formula 3.43, Johnson 1967). This indicates small distortion when the rank order of original distances is close to the rank order of ultrametrics. The correlations, no matter which formula was used, cannot be tested for significance by conventional statistical procedures, because the distance values within each matrix are not independent and, of course, the ultrametric matrix is not independent from the distances either (cf. Chapter 9). Furthermore, the correlations lack absolute validity: a value of 0.8 does not imply the same relationship between distances and dendrograms for different data structures. Additional approaches to evaluating distortion are discussed in Gower & Banfield (1975), and Gordon (1987).

5.5.2 Stability and validity

A fundamental requirement in data exploration is that the results, such as hierarchical classifications, should not change much upon a very slight modification of the starting data. That is, changes in the results should be proportional with the underlying changes of the data (*stability*). When stability is low, the *validity* of the whole classification will also be in doubt. The reverse is not necessarily true, however, because a very stable result does not mean automatically that the classification, or any part therein, provides a meaningful summary of the data.

There are several ways of evaluating stability of classifications; some emphasizing the mathematical side whereas others giving priority to the biological aspects. For example, we may examine whether random perturbations of the raw data or the distances will cause substantial changes of the resulting hierarchy (e.g., Rand 1971). The effect of deleting a variable or adding a new variable to the data upon the classification may also be examined (e.g., Jambu & Lebeaux 1983). The strategy proposed by Smith & Dubes (1980) to evaluate stability is fairly complex: the set of objects is subdivided randomly, and both subsets are subjected to cluster analysis. Two formulae have been designed to measure the extent to which the objects are grouped together in the complete classification, relative to their grouping tendency observed in the subsets. In the widest sense, analyses of the same data with different methods have also to do with stability, although in this case it is the methods, rather than the data, that change. Nevertheless, if different methods provide similar results the classification may be considered stable (as is the four-group partition of points in example **b**, see Figs. 5.6b, 5.7b, 5.11b and 5.12b). In rare instances, similarity of results is easily assessed by eye, but in general some quantitative procedure is needed to express dendrogram similarity objectively (see Chapter 9).

The fullest account of the biological relevance of stability studies is given by Rohlf & Sokal (1981a). It is a crucial issue in systematics whether a taxonomic classification changes considerably when one set of characters is replaced by another. For example, entomological studies may be based on the morphological characters of the larval stage or on the properties of the adults for the same set of OTUs. In this way one could test the *non-specificity* hypothe-

sis (Sneath & Sokal 1973: 97) which declares that there are no distinct groups of genes that affect exclusively a certain group of characters, either larval or adult. (In many studies this hypothesis has proved to be false since then, but this need not concern us at the moment.) Analogous problems in ecology arise when sites are classified based on their species composition, as well as on environmental measurements. This is in fact a little more than simple stability: predictability of one classification by another becomes the central question. The surveyor may also be interested in examining the changes of classifications when we switch from the species level to genera, families or orders (see Podani 1986, and references therein).

As we have seen, the stability of classifications is a complex matter, but validity is even more difficult to treat appropriately. Gordon (1998) has a very recent overview of the topic, by distinguishing among several objectives of such studies. In any case, the validity of a classification can only be judged by comparison to some reference basis. In external tests of validity, information not present in the starting data is used, whereas internal tests refer to the original data in the evaluation. In biology, we are usually faced with the second situation. The tests may aim at examining whether any group structure is present, and if a given cluster, a partition or a full hierarchical classification is valid. Testing group structure or the existence of a cluster involves measuring deviation from a null model, for example, the Poisson or unimodal distribution of points in the multidimensional space. The problem of validating a hierarchy has been little investigated, and the majority of methods used are those discussed later, in Chapter 9. Exceptions are techniques which follow right below.

5.5.3 The optimum number of clusters

The majority of partitioning methods require *a priori* specification of the number of clusters, so these methods should not be used unless the investigator has gained preliminary experience through some other analyses of the data. We expect, however, that hierarchical clustering, being uninfluenced by such arbitrary decisions, will be informative as to the hierarchical level where the dendrogram may be cut into an optimal partition of objects. The sample analyses illustrated earlier in this chapter led to the conclusion that dendrogram shape and the differences between subsequent levels may be more misleading than helpful, especially for the novice. Experience and knowledge of the algorithmic properties of methods are of course very useful in interpretations, but no data analyst can ever be sure about his statements made only on visual basis. Some objective methods are sought to detect the optimal level, if it exists, in hierarchical classifications. Mathematicians, biologists and others involved in classification have long been preoccupied by this problem (see Dale 1988). Milligan & Cooper (1985) summarize 30 different criteria proposed for the purpose and, understandably, their overview has become incomplete since its publication date. Most of the formulae proposed, as one could guess, incorporate computation of sum of squares, thus they are closely related to classical biometric approaches. Milligan & Cooper used artificial data sets with known properties to measure the accuracy of the different criteria in revealing the optimum number of clusters. They found that the formula suggested by Calinski & Harabasz (1974) performed the best. Let SSQ_t denote the *total* sum of squares of the data set containing n variables and m objects, calculated according to the formula:

$$SSQ_t = \sum_{i=1}^{n} \sum_{j=1}^{m} (x_{ij} - \bar{x}_i)^2 \qquad (5.13)$$

where $\bar{x}_i$ is the mean for variable *i.* (The reader may recall that Formula 3.105 defines the same quantity for a given cluster *A*). The sum of within-cluster sum of squares has also been given already, in the context of *k*-means clustering (*J*, Formula 4.1), and is now denoted by SSQ_w. The difference $SSQ_b = SSQ_t - SSQ_w$ is the part of the total sum of squares explaining between-cluster differences. The higher the between-cluster sum of squares relative to the within-cluster sum of squares, the more clear-cut the partition. Then, the idea that comes first to our mind is to examine the ratio of these two quantities, i.e., SSQ_b and SSQ_w. However, this ratio, although dimensionless, does not allow comparison of situations for different values of *k*, without division by the corresponding degrees of freedom. The criterion developed by Calinski & Harabasz will then have the following form:

$$CALHAR_k = \frac{SSQ_b}{(k-1)} \Big/ \frac{SSQ_w}{(m-k)} \qquad (5.14)$$

According to the authors, the monotone increase of *CALHAR* over *k* indicates lack of any group structure, whereas the monotone decrease suggests hierarchical relationships. If there is a maximum, the partition of objects into the respective *k* clusters can be considered the optimum(in other words, upon varying the value of *k* first the missing group structure is indicated then, beyond the maximum, the hierarchical relationships become apparent).

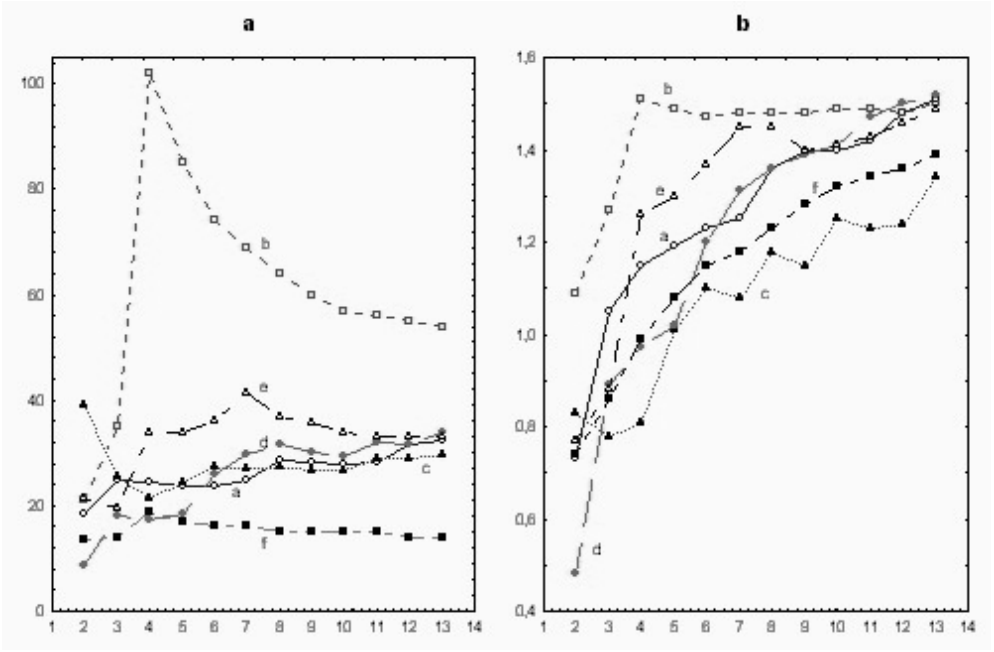

Figure 5.17. Finding the optimum number of clusters in complete linkage dendrograms of sample data sets shown in Figure 4.3. **a**: Calinski-Harabasz index, **b:** ordinal approach suggested by the author.

The above criterion can also be defined in terms of matrices defined at the end of Subsection 4.1.2. Thus, if **B** is the *nxn* matrix of between-cluster sum of squared deviations, and **W** is the matrix of within-cluster sum of squares, the criterion is expressed using the diagonal elements

$$CALHAR_k = \frac{\mathrm{tr}(\mathbf{B})}{(k-1)} \Big/ \frac{\mathrm{tr}(\mathbf{W})}{(m-k)} \qquad (5.15)$$

(see Appendix C). Let us examine the "behavior" of the Calinski - Harabasz index for increasing values of k in the complete linkage clustering results (Fig. 5.7) of sample data of Figure 4.3. The number of clusters to be considered ranges from 2 to 12, further refinement of the classification would convey very little meaning (see Fig. 5.17a). For the random **a** case the curve of *CALHAR* is definitely, although not monotonically increasing, as otherwise expected for data without having any group structure. In example **f**, which also lacks clusters, there is very little change over k, with a slight maximum at $k = 4$, showing that obscure peaks of the curve must always be treated with caution. In example **b**, the result is straightforward, with a striking peak at $k = 4$. The shape of the curve is very peculiar for the two adjacent clusters of case **c**. The start is an obvious maximum at $k = 2$, followed by an abrupt decline and then oscillation around a very low value. That the optimum partition of the points is at $k = 2$ seems acceptable, and any further subdivision of these clusters has very detrimental effects as to the "goodness" of partition. Since the elongated and arched point clouds (cases **d** and **e**) were not recognized by complete linkage analysis, we are not surprised that the maxima occur at $k = 7$ and $k = 8$, respectively, that is, for partitions that are refined enough to cause no conflicts with the non-spherical classes.

Without going into much more details regarding the methods discussed (and not discussed) by Milligan & Cooper's (1985) review, it is worthwhile to point out that almost all have geometric or at least traditional statistical interpretation. Their application seems to be restricted to situations when the sum of squared deviations is meaningful. In many clustering studies, however, the dissimilarity measure applied has nothing to do with sum of squares, so that the Calinski-Harabasz-index would be misused. A more general procedure is sought, which emphasizes compatibility with the dissimilarity function rather than the geometric interpretability. Such a procedure was proposed by the author (Podani 1998) and will now be described briefly. The basic idea is that we determine the extent to which each variable explains the given partition into p clusters. The ranking procedure may find some variables important in explaining the group structure which had otherwise very little effect upon the classificatory process. An intutitive advantage of the procedure is perhaps the implied *majority rule:* the more variables support a given partition, the more acceptable it is.

In the first step of the analysis, the absolute contribution of each variable to the within-cluster and between-cluster distances or dissimilarities is measured for, say, p clusters. The calculation of this contribution varies with the dissimilarity coefficient (see Table 5.3). For example, in case of the Euclidean distance the contribution of variable i to the distance between objects j and k is proportional to $(x_{ij} - x_{ik})^2$. Let us denote this quantity by g_{ijk}. In the next step, we examine how these contributions are distributed within- and between-clusters. Variable i fully explains the partition if all within-cluster g_{ijk} values are smaller than the between-cluster contributions; that is, based on the rank order of g_{ijk} values the sum of within-cluster ranks, R_w is the minimum. The variable may be completely neutral, so that the within- and between-cluster contributions are approximately randomly ranked (a random rank model is thus the reference). There is a slight chance that a given variable is contradictory with the partition, so that within-cluster g_{ijk} scores will have higher ranks than the between-cluster ones. This information is summarized by a single number, the ψ_{ip} criterion, given by the formula

Table 5.3. Contribution of variable i to d_{jk} for several, well-known dissimilarity and distance functions. For presence/absence coefficients, $x_{ij} = 1$ or $x_{ij} = 0$. Contributions are minimized, except for coefficients marked by an asterisk, for which contributions are maximized. n = number of variables (after Podani 1998). Numbers in brackets refer to equations of original indices, rather than the complements.

Euclidean distance (3.47)	$(x_{ij} - x_{ik})^2$
Manhattan (3.48), 1-simple matching coeff. (3.6)	$\lvert x_{ij} - x_{ik} \rvert$
Penrose *SIZE* (3.93)	$x_{ij} - x_{ik}$
Chord distance (3.54) *	$\dfrac{x_{ij}\, x_{ik}}{\sqrt{\sum_{h=1}^{n} x_{hj}^2 \sum_{h=1}^{n} x_{hk}^2}}$
Canberra metric (3.52)	$\dfrac{\lvert x_{ij} - x_{ik} \rvert}{\lvert x_{ij} \rvert + \lvert x_{ik} \rvert}$
Percentage difference (3.58), 1-Sorensen (3.25)	$\dfrac{\lvert x_{ij} - x_{ik} \rvert}{\sum_{h=1}^{n} x_{hj} + x_{hk}}$
Marczewski - Steinhaus (3.60), 1-Jacc. (3.24)	$\dfrac{\lvert x_{ij} - x_{ik} \rvert}{\sum_{h=1}^{n} \max[x_{hj}, x_{hk}]}$
1-Similarity ratio (3.71) *	$\dfrac{x_{ij}\, x_{ik}}{\sum_h x_{hj}^2 + \sum_h x_{hk}^2 + \sum_h x_{hj} x_{hk}}$
1-Russell - Rao (3.23)	$(1 - x_{ij} x_{ik}) \,/\, n$
1-Rogers - Tanimoto (3.9)	$\dfrac{2 \lvert x_{ij} - x_{ik} \rvert}{n + \sum_{h=1}^{n} 2 \lvert x_{hj} - x_{hk} \rvert}$
1-Sokal - Sneath (3.11)	$\dfrac{2 \lvert x_{ij} - x_{ik} \rvert}{\sum_{h=1}^{n} \max[x_{hj}, x_{hk}] + \lvert x_{hj} - x_{hk} \rvert}$
1-Anderberg 1 (3.12) *	$\dfrac{x_{ij} x_{ik}}{\sum_h x_{hj}} \cdot \dfrac{x_{ij} x_{ik}}{\sum_h x_{hk}} \cdot \dfrac{(1 - x_{ij})(1 - x_{ik})}{n - \sum_h x_{hj}} \cdot \dfrac{(1 - x_{ij})(1 - x_{ik})}{n - \sum_h x_{hk}}$
1-Kulczynski (3.62) *	$\dfrac{\min[x_{ij}, x_{ik}]}{\sum_h x_{hj}} + \dfrac{\min[x_{ij}, x_{ik}]}{\sum_h x_{hk}}$

$$\psi_{ip} = 1.0 - \frac{R_w - R_{\min}}{R_{\max} - R_{\min}} \qquad (5.16)$$

which takes the value of 1 for perfect explanation. $\psi_{ip} = 0$, if the variable is indifferent, and $\psi_{ip} < 0$, when the variable is contradictory. R_{min} and R_{max} are the possible minimum and maximum, respectively, of the sum of within-cluster ranks. Using the ψ_{ip} values the variables may be ranked, showing their utility in explaining the partition into p clusters.

What we have said up to this point is valid for non-hierarchical classifications, so it is time to turn back to our current problem, the evaluation of hierarchical classifications. We postulate that in a given sequence of partitions, represented by a dendrogram, the optimum is the one supported by the majority of variables. This is indicated simply by the sum of ψ_{ip} scores, which is denoted by Ψ_p. Its upper bound is obviously n, the number of variables in the data set. The change of Ψ_p in the function of p is easily illustrated in a diagram, and the shape of the curve informs us about the overall *explanatory power* of the variables. Although the method assumes an ordinal, rather than a geometric approach, and is efficient only for many variables, it is compared with the peformance of the Calinski-Harabasz-index (Fig. 5.17b). The two diverging methods produce similar results in case **b**, although the peak is less emphasized by the ordinal method. In examples **a** and **f** the curve is monotonically increasing, which is an indication of non-classifiability. The same is observed for example **d**, showing that complete linkage results do not reflect the "real" data structure. For case **e** the maximum is at p = 7 for both methods. Example **c** demonstrates by its lacking peak that the method is less efficient for very few variables, and in this case the Calinski-Harabasz-index performed much better.

5.6 Literature overview

Classification, with emphasis on hierarchical approaches, is popular in biology and related disciplines, as well as in humanities (e.g., psychology, sociology), and lately the topic has received increased attention by mathematicians as well (cf. Mirkin 1996). This is demonstrated by the mere existence of classification societies established in many countries, and by their worldwide superorganization, the *International Federation of Classification Societies*. Its official organ, the *Journal of Classification* has been published since 1984, indicating continuing popularity of the subject. The literature of classification is immense; Blashfield & Aldenderfer (1978) were perhaps the last to give a full overview to that date, since then no attempt has been made to cover the whole area. It is therefore essential to give some selected sources of information for the future orientation for anyone interested in particular details or historical facts.

The mathematical aspects of classification were first covered in review papers (e.g., Cormack 1971, Williams 1971); but soon after Jardine & Sibson (1971), Anderberg (1973), Everitt (1974), and Clifford & Stephenson (1975) provided milestone books which are still useful references. Everitt's work has reached its third edition, for example. Further important texts include Späth (1980), Gordon (1981), Aldenderfer & Blashfield (1984), Romesburg (1984), and Jambu & Lebeaux (1983). It is striking that the literature of multivariate data analysis treats classification as an insignificant, peripheral area (e.g., Mardia et al. 1979) or ignores the topic completely. Basic features of hierarchical clustering procedures are evaluated in detail by Diday & Simon (1976), Dubes & Jain (1976, 1979), Murtagh (1983), Day & Edelsbrunner (1984), Gordon (1987), and Milligan (1989). Everitt's (1979) problem-oriented summary is still thought-provoking. More recent developments have been summarized by Gordon (1996).

Two branches of biology seem to be "classification-maniac", taxonomy and phytosociology (coenology), in which the final objective of the survey is often to give a classificatory summarization of the data. If supplemented with cluster analysis, these fields can be labeled as numerical taxonomy and numerical syntaxonomy, the adjective referring to the fact that other types of methods are also useful tools (e.g., ordination). Numerical taxonomy has its origins in its first general summary, Sokal & Sneath (1963), whose second edition is an ample source of information for any biologist interested in data exploration (Sneath & Sokal 1973). Unfortunately, the book has no more editions, although major developments of the subject are covered in reviews by the same authors (e.g., Sokal 1986). Since the basic aim of numerical taxonomy is the exploration of phenetic similarity, without explicit reference to phylogeny, almost all of its meth-

Table 5.4. Appearance of main methods of hierarchical clustering in selected computer program packages developed for personal computers.

Method	**Statistica**	**BMDP**	**NT-SYS**	**SYN-TAX**	**NuCoSA**
Single/complete link	++	++	++	++	++
Average link	++	++	++	++	++
Centroid	++	++	++	++	++
β-Flexible			++	++	++
Incr. sum of squares	++			++	++
Information theory meth.				++	
Global optimization				++	
Monothetic divisive cl.				++	
Minimum spanning tree			++	++	
Cophenetic correlation			++	++	
Optimum no. of clusters				++	
Dendrogram graphics	++	++	++	++	++

ods were repressed when character-based and especially molecular cladistics (next chapter) gained popularity in systematics. Although Stuessy (1990) considers them as being equally important in plant systematics, Mayr & Ashlock (1991) take the view that numerical taxonomic methods that are not suitable for phylogenetic reconstruction have very little practical and theoretical importance. I think that the paradigmatic shift from numerical taxonomy to cladistics does not render hierarchical clustering methods completely out-dated in systematics, especially in studies below the species-level. We return to these problems in the subsequent chapter. Other books treating numerical taxonomy at the introductory level are Cole (1969) and Dunn & Everitt (1982). Pankhurst (1991) examines relationships between principles of classification methods, data bases and identification keys. Sneath (1995) summarizes the more than 30 years of history of numerical taxonomy, emphasizing its decisive role in the establishment and development of modern approaches to taxonomic data analysis, namely cladistics and the new morphometry (the latter discussed in Chapter 7).

Classification problems of syntaxonomy and general ecology have been treated to considerable extent in a wide variety of books. Here the fundamental controversy has appeared to be for a long time between classification and ordination, although the debate was never as futile as the phenetics vs cladistics opposition. There are several works equally important for classification and ordination theory and applications. Whittaker (1973), Williams (1976), Orlóci (1978), Gauch (1982), Greig-Smith (1983), Kershaw & Looney (1985), Legendre & Legendre (1983), Digby & Kempton (1987), Jongman et al. (1987), Ludwig & Reynolds (1988) és Kent & Coker (1992) are just arbitrarily chosen examples from the vast literature. Green (1979) is recommended to those wishing to classify both biological and environmental data, whereas Klijn (1994) provides a starting point for ecosystem-level classifications. Information theory clustering is summarized in Feoli et al. (1984). Two proceedings merit further attention of vegetation scientists: Mucina & Dale (1989) and Feoli & Orlóci (1991). Of the review articles Maarel (1979) and Gauch & Whittaker (1981) are somewhat old, for more recent accounts see Kent & Ballard (1988) and Mucina (1997).

5.6.1 Computer programs

Methods of hierarchical clustering are found in several packages, although there is a very narrow choice among options, restricted usually to a few of most widely-known algorithms. Table

5.4 lists program packages that provide an acceptable user-friendly working environment on PCs. These programs lack several special methods that may attract the biologists' attention. In such circumstances, the reader should consult books and papers supplemented with program lists, for example, Anderberg (1973), Späth (1980), and Orlóci (1978), although adaptation of the listed programs to our machine may turn out to be a tiresome task. To mention only two particular examples, the **CONISS** program (list in Grimm 1987) is the constrained version of the incremental sum of squares method, whereas the list of **TWINSPAN** appears in Hill (1979a). Blashfield (1976) has a summary of software developed in the golden age of mainframe computers, but this enumeration has become largely obsolete.

5.7 Imaginary dialogue

Q: *The examples seem to suggest to me that clustering methods by themselves very often fail to give a meaningful representation of the data, and some are the least applicable to the very purpose of the analysis, namely classification!*

A: Your criticism is very serious, and it is true that we are faced with a paradoxical situation. No computerized clustering procedure can be recommended as the only and universal classificatory recipe, superseding the other possibilities. Different clustering methods, as alternative tools of data exploration, can provide diverging results, each contributing just a little to a final picture which is called the classification. Contrary to expectations of people twenty or more years ago, clustering methods cannot replace completely the personal, yet cautious and careful judgement of the investigator (Dunn & Everitt 1982). A taxonomist cannot be satisfied with a single input of his data set into the black box of the computer and cannot expect that the output will be an absolutely objective and irrefutable classification of his study organisms. The time when journals were more than happy to publish articles abounding in uncritical applications of clustering is over. In the complex and artistic process of classification, other methods not designed explicitly to detect group structure in the data (e.g., ordination and seriation) are as important as the clustering algorithms themselves. These additional tools of the classifier will be introduced to you in the subsequent chapters.

Q: *There were big differences among alternative results of different methods, and the resulting hierarchies were occasionally so "artificial" that I am tempted to discard hierarchical methods almost completely from the arsenal of data analysis.*

A: No, do not be so pessimistic and do not give it up so quickly! You can pretty well have similar impression after reading the forthcoming discussion of the other methods! A fact that cannot be emphasized strongly enough is that there are many methods available and there are even more possible data structures, and these two things never meet automatically.

Q: *As a matter of fact, is hierarchy so important by itself? Have you not forgotten about this problem when you evaluated the sample dendrograms by only checking if hierarchical clustering is suitable to detect certain non-hierarchical group structures of the data? Do not you feel some contradiction here?*

A: It is a good point again. I did not show you any example in which a full hierarchy, equally from the top to the very bottom, would have been of our concern. Perhaps, such an example is less illustrative, by adding very little to the understanding of the problem and, believe me: if

clear-cut groupings appear at several levels, then any method, which recognizes clusters at one level, will detect them at the other levels as well.

Q: *Yes, I remember your statement that in constrained classification the full hierarchy is usually less interesting than a particular partition.*

A: And this is the case with many other clustering methods, because full hierarchies, resolved down to the individual objects are rarely needed in practice (exceptions common in numerical taxonomy). Almost all hierarchical methods have their non-hierarchical counterpart which can be used to improve a partition obtained by cutting the dendrogram at a given level. The *k*-means procedure can be a good supplement to all hierarchical methods that optimize the sum of squares of clusters, for example. *A posteriori* iterations may not always be necessary, however, but in most cases "dendrogram cuts" can be improved by the subsequent partitioning. But be careful: the hierarchical and non-hierarchical methods should be compatible with one another – otherwise you end up in a methodological confusion.

Q: *We know that overlapping partitions may sometimes give a more faithful representation of data than hard ones. What about overlapping hierarchical classifications?*

A: There have been some attempts to express uncertainties of hierarchical classifications in the dendrograms (e.g., in the fungus classifications of Dabinett & Wellman 1978) which appear analogous to overlapping partitions. The overlaps were expressed by line segments connecting different branches of the tree, which become very confusing if the degree of overlap is high. As far as I know, overlapping hierarchies have been completely neglected in recent studies.

Q: *Then, let me continue with guesses: by the same token fuzzy hierarchies could also be conceived!*

A: Yes, fuzzy "dendrograms" do exist. Marsili-Libelli (1989) proposed to construct dendrograms based on the maximum weights for each object, but this is not free from reversals.

Q: *I feel that large amounts of data can cause computational problems, as in the case of non-hierarchical clustering. The algorithms following the strategy of Fig. 5.3b seem to be the most economical, as far as memory requirement is concerned, but even the stored matrix approach seems to be impractical for several thousands of objects.*

A: Yes, you are right, but I can assure you that there have been proposals for rapid hierarchical clustering of large amounts of data. The divisive strategy of Jambu (1981, with program list) can handle up to 5000 objects, but the analysis stops after the fiftieth division, so that we do not get a complete hierarchy. As I said, you are usually not interested in the finer details of hierarchical classifications, so the upper 50 levels are sufficient. Experiencing the fast development of computer technology, however, I assure you that 5000 objects will not cause problems for PCs in the nearest future.

Q: *Yet another provocative question, but I can anticipate your response: how about classificatory space series?*

A: Of course, there are series defined in the classification space, exemplified already by the flexible strategy (Fig. 5.10). Generating such series is fairly easy by systematically changing the appropriate parameters of flexible methods. You can find further examples in Podani (1989b).

Q: *The identification key presented at the end of Chapter 3 is certainly most useful to assist the novice to navigate in the complex area of dissimilarity functions. Could you please provide me an analogous key to the hierarchical algorithms?*

A: Although such a key is badly needed, I do not think it is possible to give one for the beginner. The choice among dissimilarity functions depends on many factors, but in a given study the range of potentially useful functions can be narrowed in a meaningful way, allowing construction of a key. This is not so with hierarchical clustering, and the best I can do is to give some advice. In a clustering study the single, complete and average link methods should be tried simultaneously, because they indicate contrasting aspects of group structure. If our data allow, i.e., the sum of squares and the averages are meaningful, then the incremental sum of squares method is also worth trying. These four methods are so widely-known that our results will be understood for other people with ease. In addition, you may use other techniques of your own choice, and the more you try the more information you extract. And a final note: do not use hierarchical clustering without supplementary ordinations!

Q: *I wonder how many ways can m objects be classified hierarchically. I can only guess that this number is much higher than the number of partitions into k groups.*

A: True, the number of partitions into k clusters (given by Formula 4.17) is insignificant compared to the number of different hierarchical classifications, for any value of k. If we consider only the topology of bifurcations and forget about hierarchical levels, then for m objects we have

$$V_m = \frac{(2m-3)!}{2^{m-2}\,(m-2)!} \tag{5.17}$$

different trees (Cavalli-Sforza & Edwards 1967, Phipps 1975). These dendrograms may be called the partially ranked trees (Podani 2000), because ordering of levels is meaningful only along a given branch. For $m = 10$, this formula yields about 34 and a half million, so you can easily appreciate that it is impossible to examine all the possible dendrograms for a much larger set of objects. If the order of levels in the whole dendrogram is deemed important (fully ranked dendrograms), then the number of possibilities will be:

$$D_m = \frac{m!\,(m-1)!}{2^{m-1}} \tag{5.18}$$

(Frank & Svensson 1981). For 10 objects, Formula 5.18 yields more than two and a half billion! Of course, when you consider the absolute values of levels, the number of possibilities is infinite.

Q: *Formula 5.18 is unbelievably simple. One would expect some more complicated equation, so I would be happy to see its derivation.*

A: Here you are. We have m objects and assume, without losing generality, that one fusion is performed in each step in agglomerative clustering analysis. So, in the first step we have $\binom{m}{2}$ pairs to choose from. In the next step, we shall have only m-1 items, that is m-2 objects and one cluster, so that the number of choices is $\binom{m-1}{2}$. The number of pairs is calculated in the same

manner until the last fusion, where we have two clusters and the trivial $\binom{2}{2} = 1$ possibility of their merger. Since the fusion in each step is independent of the fusions of the other steps, these terms can be multiplied to get the number of different dendrograms, that is

$$\binom{m}{2} \times \binom{m-1}{2} \times \binom{m-2}{2} \times \ldots \times \binom{3}{2} \times \binom{2}{2} \tag{5.19}$$

Rewriting this we obtain

$$\frac{m(m-1)}{2} \times \frac{(m-1)(m-2)}{2} \times \frac{(m-2)(m-3)}{2} \times \ldots \times \frac{3 \times 2}{2} \times \frac{2}{2} \tag{5.20}$$

which, after simplifying the multiplication terms, reduces to Formula 5.18 immediately. In this derivation it is implied that ordering of levels is important, otherwise fusion of a pair, say 1-2, in the first step and the fusion of another pair, say 3-4, in the next step could be interchanged without affecting the result. If this interchangeability is allowed, Formula 5.17 specifies the number of dendrograms.