9

Comparative evaluation of results

(The analyses must go on!)

Classifications, ordinations and other types of results do not necessarily represent the final stage of our endeavour into the realm of multivariate data exploration. There are several arguments supporting the view that the computations should go on in most cases. For example, it is mostly true, except for trivial cases, that different procedures applied to the same set of data produce more or less diverging results! It was demonstrated in this book most markedly for hierarchical clustering. Although the methods themselves are considered 'objective' tools, there are several points during data analysis where the surveyors' decisions are inevitably subjective. To mention a few: the definition of sampling characteristics (e.g., quadrat size), the selection of variables, data types and transformation methods, the choice of the resemblance function, the ordination or clustering algorithm are all up to the investigator, – and the list could have been continued. To make sure that these decisions do not influence our conclusions significantly, it is always advisable to examine their relative impact upon the results. This is the only possibility to remove the methodological 'artefacts' from the analysis, thus revealing information that truly reflects the properties of study objects themselves. The comparison of the results of alternative analyses is the most useful in this approach. In certain situations, however, comparisons do not relate to methodological choices. A good example is parsimony analysis in cladistics which may produce hundreds of equally optimal trees whose synthesis into a new tree leads us to the final conclusions. This chapter is devoted entirely to various approaches to the comparison and synthesis of alternative results.

Units of comparison

Each alternative result may be considered as an object, in the same way as taxa, sample plots and other individuals were treated in the first part of the study. On the analogue of OTUs of numerical taxonomy and OGUs of geography, dendrograms, ordinations, distance matrices and other types of results may be collectively termed as *operational units of comparison* (OUC, Podani 1989d). Whereas two dendrograms may be contrasted according to the usual

Euclidean distance function in the same way as two OTUs, the characteristics on which the comparison is based are very different from the features of 'natural' objects. It is clear for all of us that the comparative evaluation of OUCs requires special means that adequately reflect their mathematical properties. When the distance is properly defined, then most analytical tools already known from the previous chapters will be helpful, although in certain situations the old and good procedures of conventional biostatistics are called for – in a modified form.

9.1 Main choices

Comparisons may follow very diverse logical pathways, therefore this topic is extremely intricate and complicated. For didactic reasons and better orientation in the subject, the main possibilities are categorized. We should note first that the investigator is faced with several choices between two alternatives when making a comparison, even though he or she is not always aware of any such decision (Podani 1989d). Many choices fit a dichotomous decision tree (Figure 9.1) whereas others – the latter three in the list that follows – have a more general validity, being equally important on several branches of the tree.

9.1.1. Type of results: identical vs different

Comparisons are most commonly made between results of the *same type*, such as between two partitions or two ordinations. This possibility was not explored yet in this book. However, we have seen already some examples of an approach in which the units compared are of *different type*. The comparison of a dendrogram with the matrix from which it was derived (cophenetic correlation, Subsection 5.5.1) implies a quantitative measurement of the agreement between two different kinds of mathematical objects. The simultaneous graphical display of two results by the superposition of one result upon the other is an example for a visual comparison, as illustrated by positioning a plexus graph over an ordination diagram (Figure 8.10b). Further examples for inter-type comparison will be shown in Subsection 9.5.2.

9.1.2. Similarity vs consensus

The similarity or distance between a *pair* of OUCs may be expressed numerically, and the comparison of $k > 2$ OUCs in all possible pairs ('multiple comparisons') provides a resemblance matrix which may be used in turn for the classification or ordination of OUCs ('meta analysis'). The very same k results may also be synthesized into a k+1th result which may show both agreements and disagreements of the original results. This synthesis is the *consensus* object often used to represent the entire set of competing OUCs in the biological interpretation of results.

9.1.3. Hypothesis testing vs exploratory analysis

The investigator may want to see whether the similarity between two OUCs is significant or not in the traditional sense. That is, the data exploration may end up with a statistical approach which was almost completely forgotten in the first phase of the study! In order to be able to do this, however, two important conditions must be satisfied. First of all, the two OUCs to be compared must be independently derived which means, for example, that they cannot come from the same set of data. Thus, the significance of dendrogram similarity may only be tested

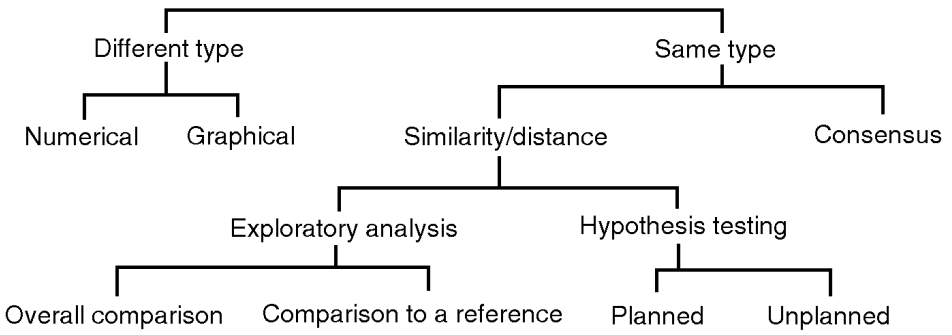

Figure 9.1. Tree diagram illustrating the main choices to be made when comparing results of multivariate analysis.

if the first is based on variable domain A and the other on variable domain B, the domains having no variables in common. The question whether the two dendrograms are significantly similar implies testing the proposition that the two groups of variables lead to similar classifications. The other condition of any significance test is the availability of the reference distribution of the statistic that measures similarity. Since the underlying distributions are not known, with a few exceptions, the only resolution is to generate them by Monte Carlo simulation. The essence of this approach is that hundreds or even thousands of randomly created pairs of OUCs are compared by the given similarity measure, and the frequency histogram of categorized similarity values is drawn, allowing to examine the position of the actual similarity value.

If the independence condition for significance tests is violated, then some *exploratory* function of the statistics still remains. For $k > 2$, we can proceed as described below in subsection 9.1.5. However, in studies restricted to the comparison of two OUCs the single similarity value is practically uninformative by itself. In such a case, we can still use the reference distribution to assess the position of the single value relative to the mean, etc., but we should never make statements as to the 'significance' of such results.

9.1.4 Planned vs unplanned comparisons

If significance testing is valid and there are several pairs of OUCs, then we must give careful considerations to the following problem. We are faced with a situation analogous to determining the least significant difference (LSD) after the ANOVA of several samples (Sokal & Rohlf 1981a): the selection of significant similarities from the matrix of all $\binom{k}{2}$ values accumulates Type I errors and therefore more pairs are deemed significantly similar than actually are at the chosen probability level (usually $p = 0.05$). Whenever we decide before the calculations that some particular pairs are of interest only (*'planned comparisons'*), then the above problem is avoided and the simulated distribution applies to the test. If, on the other hand, there are no *a priori* selected pairs of OUCs (*'unplanned comparisons')* and all pairs are to be tested, then the test should be made more conservative. The Monte Carlo simulation of the minima for $k(k-1)$ pairwise comparisons provides a solution for this problem (see Subsection 9.3.6).

9.1.5. Overall comparisons vs comparisons to a reference basis

We do *overall* comparisons when none of the OUCs is favored for some reason, so that comparisons in all possible pairs are plausible. The relative differences between the similarity values will be most informative in the meta-analysis that follows. No such meta analysis is required, however, if one of the OUCs is considered as a *reference basis* to which all the others may be compared. For example, an ordination based on all the variables is the reference and we may wish to examine how the stepwise omission of least important variables will modify similarity to this reference ordination. The reference now serves as the 'control' object in such comparisons.

9.1.6. Congruence vs algorithmic effects

Rohlf & Sokal (1981b) and Gower (1983) called our attention to this distinction, not shown in the decision tree of Fig. 9.1 because it appears logically in all comparisons. This is essentially a distinction between theoretical/biological reasons and technical/methodological aspects. We can say that if the difference between alternative results may be explained by biological causes, then we do analysis of *congruence* (e.g., comparison of classifications based on data from the adult and larval stages to evaluate *taxonomic congruence*). Such problems are better distingushed from situations when the differences among results are explained by mere *algorithmic* modifications and other technicalities.

9.1.7. Elementary vs complex comparisons

In *elementary comparisons*, the differences between the alternative OUCs are caused by a single factor. In a study of the effect of classificatory strategies upon the dendrograms, the other 'parameters' of the analysis (data type, resemblance function, etc.) must be kept constant. If we do not care about this, then the change of two or more factors will have a confounding effect upon the results, and our conclusions may be misleading (cf. Kenkel & Orlóci 1986). Such confounding effects are disregarded more commonly in the published literature than we would think! If two or more factors are evaluated systematically, in all possible combinations, then their relative importance may be revealed by *complex comparisons* (Podani 1989d).

9.1.8. Uni- vs multivariate evaluation

Any comparison is *univariate* when a single property of the OUCs is considered (e.g., contrasting dendrograms with one another based on path differences only, see Subsection 9.2.3). Quite surprisingly, most of the published comparisons are of this type, even though all the previous steps of the analysis are essentially multivariate in nature! Logic dictates that the full chain of computations should be multivariate, wherever possible. Podani & Dickinson (1984) argued that, in case of dendrograms at least, the comparisons may be based simultaneously on several properties of results, so that the entire study may be multivariate. This is also possible for other, relatively complex objects such as cladograms, additive trees and minimum spanning trees.

9.2 Pairwise comparison of results

Most of the relevant studies incorporate pairwise comparisons, which may be of central importance even in consensus generation. Therefore, the methodological part begins with this subject although the decision between pairwise comparison and consensus is not on the highest level of the hierarchy of Figure 9.1. The methods suitable for comparisons are discussed separately for each type of result. First, procedures for the comparison of resemblance matrices are introduced because the evaluation of other types of results may often be traced back to matrix comparisons.

9.2.1 Matrix comparisons

Numerical methods. Two symmetric resemblance (distance, similarity, etc.) matrices, denoted here by **D** and **E**, are prepared for numerical comparison by unfolding each of their upper semimatrices into a column vector. Then, we can make a choice from the huge arsenal of resemblance functions discussed in Chapter 3. Most often, the correlation coefficient (Formula 3.70) is adapted, best known under the term *matrix correlation* (Sneath & Sokal 1973: 280):

$$r_{\mathrm{DE}} = \frac{\sum_{i=1}^{m-1} \sum_{j=i+1}^{m} (d_{ij} - \bar{d})(e_{ij} - \bar{e})}{\left[\sum_{i=1}^{m-1} \sum_{j=i+1}^{m} (d_{ij} - \bar{d})^2 \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} (e_{ij} - \bar{e})^2 \right]^{0,5}} . \qquad (9.1)$$

In this, $\bar{d}$ and $\bar{e}$ are mean resemblance values for **D** and **E**, respectively. The values in the diagonal are excluded from the averaging and from the comparison. If the two matrices imply similar tendencies for the resemblance/distance relationships of objects, irrespective of the absolute magnitude of values, then their matrix correlation will be close to 1. In general, r_{DE} falls into the interval [–1, 1], as usual for correlation measures. It must be pointed out that the application of correlation in this case is exploratory, rather than rigorously statistical, because the values within each matrix are not independent of one another. Thus, the 'significance' of r_{DE} cannot be tested in the usual manner (see Subsection 9.3.1, for more). The Euclidean distance between **D** and **E** may also be calculated according to:

$$d_{\mathrm{DE}} = \left(\sum_{i=1}^{m-1} \sum_{j=i+1}^{m} (d_{ij} - e_{ij})^2 \right)^{1/2} . \qquad (9.2)$$

Other formulations also appear in the literature, but these two functions are the most common in practice. Matrix correlation is most informative together with graphical comparisons (see below) and when the two matrices are not commensurable (one is distance and the other is dissimilarity, for example). For meta-analysis, all pairs of OUCs are compared by the complement of correlation, as illustrated in the following example. Euclidean distance is meaningful only if the two resemblance matrices have identical measurement scales.

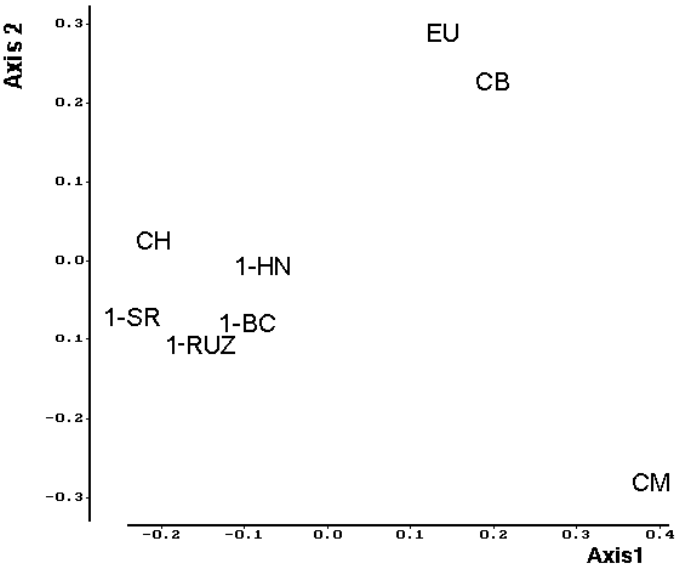

Figure 9.2. Principal coordinates ordination of eight resemblance matrices obtained by different dissimilarity measures.
Symbols:
1–BC: Bray - Curtis,
1–RUZ: Ruzicka,
1–SR:similarity ratio,
1–HN: Horn index,
CM: Manhattan metric,
CB: Canberra metric,
EU: Euclidean distance,
CH: chord distance.

The dissimilarities among the objects of Table A1 are calculated first, based on eight coefficients. The resulting matrices are compared in all possible pairs using the complement of Formula 9.1. The advantage of using correlation in this case is that the eight measures express dissimilarity on different scales. The matrix of matrices is then evaluated by principal coordinates analysis (Figure 9.2). The first two axes account for 84% of the distance relations, which is a relatively high percentage. The diagram illustrates perceptively the relationships of the selected resemblance measures in this study. The five-member group on the left side comprises measures of very similar behavior; and the Euclidean distance and the Manhattan metric also form a small group. The latter is perhaps surprising because Euclidean distance emphasizes squared differences, rather than absolute deviations as in the Manhattan metric. The Canberra metric has an odd performance, owing to the separate standardization for each pair of variables. Similar results may be obtained easily for other data sets (as in Podani 1994: 191), showing the fair generality of the present conclusions.

Graphical procedure. In an orthogonal coordinate system, each point represents object pair *jk* with coordinate d_{jk} on the horizintal axis and e_{jk} on the vertical axis. The scatter diagram thus obtained (*matrix plot*, Rohlf 1993a) is interpreted similarly to the Shepard-diagram (Subsection 7.4.2). The better the fit of points to an imaginary line in the plot, the higher is the similarity (linear correlation) of the two matrices being compared.

The graphical comparison of matrices is illustrated for two pairs of matrices taken from the example on the previous page: 1–BC vs 1–RUZ (r = 0.994) and CM vs CH (r = 0.522). The scattergrams are shown in Figure 9.3. For the first pair, the relationship is almost linear, whereas for the second pair the similarity is much weaker. Note, for example, that the object pair with the smallest chord distance is very distant if compared according to the Canberra metric.

9.2.2 Comparison of partitions

Some methods use the strategy of matrix correlation, while others start from cross-partitions. Further, more specialized techniques evaluate the possibilities of transforming one partition into the other in order to derive their similarity.

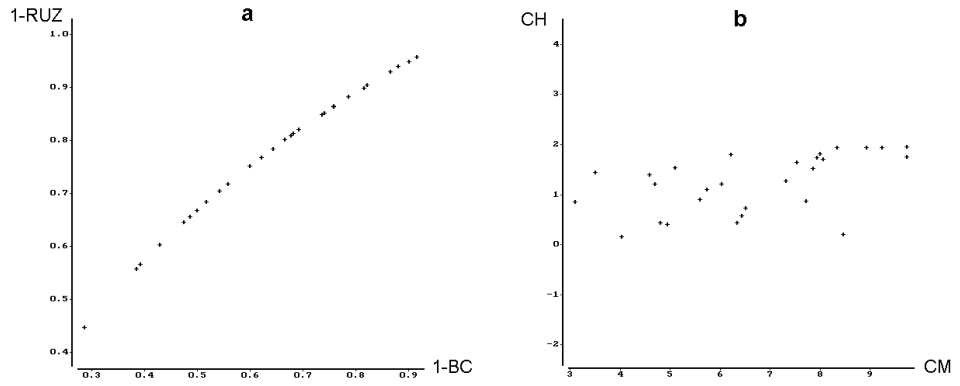

Figure 9.3. Graphical comparison of matrices. **a:** 1–Bray-Curtis vs 1–Ruzicka, **b**: Canberra metric (CM) vs chord distance (CH), each calculated for the objects of Table A1.

Matrix comparisons. Any partition P can be described in terms of an $m \backslash m$ symmetric incidence matrix, denoted here by $\mathbf{C}_P$. In this matrix, $c_{gh} = 1$ if objects g and h belong to the same class in P, otherwise $c_{gh} = 0$. Then, the similarity (or dissimilarity) of partitions P and Q is expressed by the matrix correlation between the respective incidence matrices $\mathbf{C}_P$ and $\mathbf{C}_Q$. The similarity of these matrices can also be calculated by practically any of the presence/absence coefficients discussed in Section 3.2. Using the notations of the 2×2 contingency table, a is the number of object pairs that appear in the same cluster in both partitions compared, b is the number of object pairs appearing together only in the first partition, and so on. The formulae are not repeated here, only a brief list is provided to show that their names may differ from those known from the literature of dissimilarity functions:

- simple matching coefficient (Formula 3.6, = "Rand" index, Rand 1971),
- Euclidean distance (Formula 3.7, = "PAIRBONDS", Arabie & Boorman 1973),
- Jaccard index (Formula 3.24, Downton & Brennan 1980),
- Sorensen index (Formula 3.25, = "percent mutual matches", Arabie & Boorman 1973) and
- Ochiai index (Formula 3.26, Fowlkes & Mallows 1980).

All these measures but Euclidean distance are expressed usually in form of their complements. Thus, complete agreement between partitions is indicated by 0 in every case. However, it is not true that maximum possible disagreement between P and Q yields a dissimilarity of 1, because the partitions are constrained to agree in some object pairs. The value of a cannot be zero because we cannot generate two partitions of m objects (except for the trivial cases) such that all object pairs occurring in the same group in P are in different groups in Q. In order to be able to compare dissimilarities coming from different circumstances, standardization is necessary. Usually, standardization is based on the expectation for randomly generated partitions and the potential maximum, according to the formula:

$$\frac{\text{Actual value - Expected value}}{\text{Possible maximum - Expected value}}, \qquad (9.3)$$

(see Hubert & Arabie 1985). Those authors pointed out that determining the maximum is a difficult problem of combinatorial optimization. Podani (1986) proposed heuristic searching methods to approximate the maxima. A partial and usually satisfactory solution is the comparison of actual values with the expectations and significant values, at a given probability level, obtained from simulations.

Cross partitions. The cross partitions well-known from block clustering are interpreted now as contingency tables in which the rows represent the classes in P, the rows correspond to the classes in Q. The size of the table is $s \times t$, with s and t as the number of clusters in P and Q, respectively. The value of cell ij in this table is the number of objects that belong to class i in P, and to class j in Q. For example, for the two partitions of 10 objects

1: { 1, 2, 3, 4, 5 } { 6, 7, 8, 9, 10 }

2: { 1, 2, 3, 6, 7 } { 4, 5, 8, 9, 10 }

the cross-partition table will be

		Partition 2	
		Class 1:	Class 2:
Partition 1	Class 1:	3	2
	Class 2:	2	3

which corresponds to the following subsets:

{ 1, 2, 3 } { 4, 5 }

{ 6, 7 } { 8, 9, 10 }

In general, the contingency table takes the following form:

		Q			
		q_1	q_j	q_t	
P	p_1				
	p_i		n_{ij}		$n_{i.}$
	p_s				
			$n_{.j}$		$n_{..} = m$

The marginal totals are cluster sizes in P and Q, m is the grand total (the number of objects classified). The table may be evaluated by the well-known χ^2 statistic (Formula 3.36) which is rewritten using the above notations as,

$$X^2 = \sum_{i=1}^{s} \sum_{j=1}^{t} \frac{(n_{ij} - n_{i.} n_{.j} / m)^2}{n_{i.} n_{.j} / m} . \qquad (9.4)$$

Zero value results of all classes in P are dispersed equally among the groups of Q, whereas the maximum occurs if P = Q. This maximum value, $m \times \min [(s-1), (t-1)]$, can be used as a normalizing constant in the same way as in the Cramér-index (3.37).

The Goodman - Kruskal (1954) lambda (3.38-3.39) is also applicable to comparing partitions, with the following interpretation. Suppose that first we wish to make a guess about the cluster membership of an object in partition Q without any information on its position in P. Clearly, the best trial is the largest group in Q, so we find $\max_j [\, n_{.j} \,]$ because this will minimize the number of bad guesses. However, if we do know that the object is classified into group i in P, then only the ith row of the cross-classification table should be examined and the highest value of this row, $\max_j [\, n_{ij} \,]$, is to be found. Then, based on our knowledge of P, the mean decrease of our uncertainty regarding the group membership in Q becomes:

$$LAS_{\mathrm{PQ}} = \frac{\sum_{i=1}^{s} \max_j [n_{ij}] - \max_j [n_{.j}]}{m - \max_j [n_{.j}]} . \tag{9.5}$$

It is an asymmeric measure of *predictability* or *predictive power*. Its value is zero if P is completely uninformative on Q and 1 if the two partitions are identical. This coefficient is useful when comparisons are made with a reference partition (recall Subsection 9.1.5). The symmetric measure of mutual predictability, the Goodman-Kruskal's lambda itself is calculated according to:

$$\Lambda_{\mathrm{PQ}} = \frac{\sum_{i=1}^{s} \max_j [n_{ij}] + \sum_{j=1}^{t} \max_i [n_{ij}] - \max_j [n_{.j}] - \max_i [n_{i.}]}{2m - \max_j [n_{.j}] - \max_i [n_{i.}]} . \tag{9.6}$$

Its values range from 0 to 1. The complement of (9.6) is a dissimilarity between P and Q.

> It is to be noted that there is a formal relationship between the cross partition-based and matrix-based comparisons: one may be expressed in terms of the other. For example, the value of a in matrix comparisons may be written using the notation of cross-partitions as $[\sum \sum n_{ij}^2 - m] / 2$.

Transformation metrics. The procedures most specific to partitions examine the number of elementary steps necessary to convert partition P into Q. The *transformation metric* proposed by Day (1981, *MINDMT, "min. divisions, mergences and transfers*") is the simplest of all: this is the minimum number of objects that must be reassigned to a different group in order to obtain partition Q. If $s = t$, then the cross-partition table may be transformed into a matrix $\mathbf{Z}$ such that the sum of the diagonal values is maximum and therefore the number of objects to be regrouped, the sum of off-diagonal values, is minimized:

$$MINDMT_{PQ} = m - \mathrm{tr}\,(\mathbf{Z}) \tag{9.7}$$

When $s \neq t$,then the above formula also applies provided that empty (dummy) classes are added to the partition with the fewer number of groups. The maximum of Formula 9.7 also deserves our attention, because it is useful to derive a normalized version of the coefficient:

$$MISC_{PQ} = \frac{m - \mathrm{tr}\,(\mathbf{Z})}{\max [MINDMT]} \tag{9.8}$$

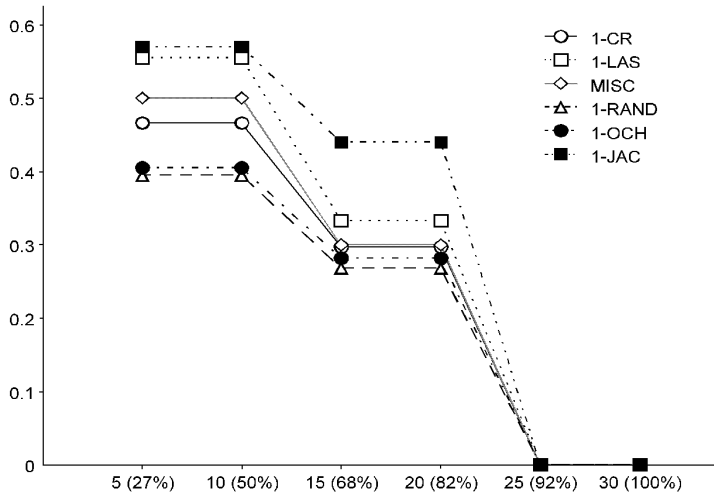

Figure 9.4. The effect of species removals on classifications visualized by the comparison to the reference (30 species, 100 %), using six coefficients of partition disagreement. The classifications were made using the presence/absence version of Table A4. The vertical axis measures dissimilarity to ther eference.

("*misclassification*" index). Its lower bound is 0 (full agreement) while the upper bound is 1 (when the maximum number of relocations are needed to convert P into Q). The maximum occurs if the values of the cross classification table are the most uniform. Day (1981) has proposed many other formulations, such as the *sigma metric*, which combines some of the indices already described:

$$\sigma_{PQ} = 2a + b + c - 2\sum_{i=1}^{s} z_{ii}^2 + 2\sum_{i=1}^{s} z_{ii} \,. \qquad (9.9)$$

This is also obtained by maximizing tr {**Z**}. The solution is not necessarily unique, however, because the same sum of squares may result for the diagonal values from different rearrangements within the cross-classification matrix (the associated *a, b* and *c* values ahve some freedom to change).

To demonstrate the above approaches, let us perform the following study. First, convert the data in Table A4 into presence/absence form and rank the 30 species according to criterion 5.8 (see Subsection 8.1.1). Then, the global optimization partitioning strategy (Subsection 4.1.2) and the simple matching coefficient (Formula 3.6) are used to classify the 20 objects into two clusters based on all species and on consecutively reduced species subsets. The reference basis is obviously the classification based on the total set of species. The dissimilarity of the other classifications to this reference may be illustrated by a line diagram (Fig. 9.4). This allows demonstrating the relationship between a classification and the number of variables considered. Many coeffcients of partition agreement will also be comparable (here I consider only those producing a range of [0,1], so the sigma metric and PAIRBONDS are omitted). The removal of the least important five species does not modify the starting classification, but leaving further five species out will be influential. Twenty and fifteen species provide the same classification, and the same is true for 10 and 5 species. The index most sensitive to the initial changes is 1–*JAC*, which increases slowly afterwards. Since the maximum number of relocations is 10, the value of the *MISC* coefficient informs us that the first big change involves the relocation of three, and then five objects from the intial groups. Actually, this index seems to reflect quite well the 'average behavior' of the other five indices.

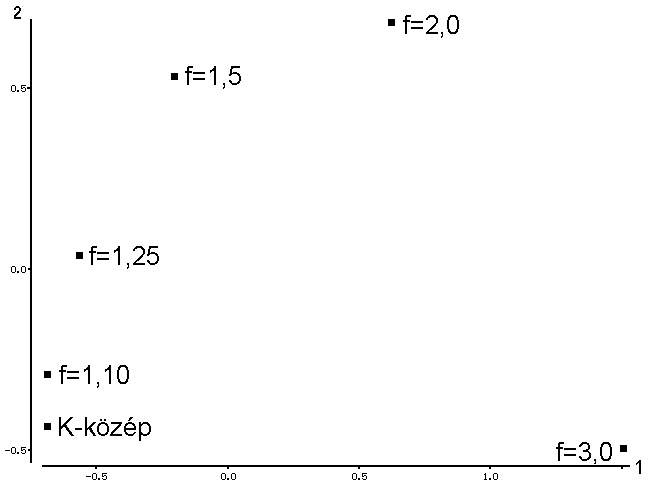

Figure 9.5. A classification series as portrayed by a PCoA ordination. The *Iris* individuals (Table A2) were classified by the *k*-means method and the fuzzy *c*-means clustering algorithm with increasing values of the coefficient of fuzziness (*f*). Note that the arch effect may also appear in meta-analysis.

Comparison of fuzzy partitions. This approach is based on the $\mathbf{U}_{m,c}$ matrices in which the weight $0 \leq u_{jk} \leq 1$ measures the degree of belonging of object j to class k. Two fuzzy partitions F and G may be adequately represented by the corresponding matrices $\mathbf{U}_F$ and $\mathbf{U}_G$. Podani (1990) suggested measuring the dissimilarity between F and G by the minimum sum of squared deviations of the values in $\mathbf{U}_F$ necessary to transform F into G. This symmetric measure is obtained by examining all the column permutations of $\mathbf{U}_F$ while $\mathbf{U}_G$ remains unchanged. More formally, we minimize the quantity

$$\Delta_{FG} = \left(\sum_{j=1}^{m} \sum_{k=1}^{c} (u_{Fjk} - u_{Gjk})^2 \right)^{1/2}, \qquad (9.10)$$

in which c is the number of classes. The permutations are easily generated up to 7 or so classes; larger values of c rearely appear anyway. The above formula tolerates unequal numbers of classes in F and G; only dummy classes are to be added to the classification with the fewer number of groups. Clearly, Formula 9.10 applies to hard partitions as well; recall that they are just special cases of hard partitions (for each object, one weight is 1 and all others are zero). There is a simple relationship between 9.10 and 9.7: $\Delta^2 = 2MINDMT$. The maximum of 9.7 is therefore useful for normalizing Formula 9.10 as well.

The method is illustrated using the *Iris* data set. The 150 individuals are assigned to three classes by the fuzzy *c*-means clustering algorithm, with the following values of the coefficient of fuzziness: 1.10, 1.25, 1.5, 2.0 and 3.0. Since the coefficient cannot attain the value of 1, for singularity problems, the *k*-means classification of the same objects is considered as a comparable hard partition. These operations provide a classification series the members of which are compared in every possible pair using Formula 9.10. The dissimilarity matrix is then analyzed by principal coordinates analysis (Figure 9.5). The first two axes account for 80% of the variation, indicating that the two-dimensional scattergram is a quite faithful representation of the relationships among the classifications. Notwithstanding the presence of an arch in the resulting configuration, there is a clear 'gradient' from the hard partitions towards the fuzziest one.

9.2.3 Dendrograms and cladograms

Being tree graphs, dendrograms and cladograms are more complex structures than the OUCs discussed thus far snd their comparative evaluation is a real challenge for us. Regarding their inherent topological structure, a dendrogram and a rooted cladogram are similar: both of them summarize hierarchical relationships in form of a usually dichotomous tree with the terminal nodes representing the objects. (I will not discuss unrooted phylogenetic trees here). In most dendrograms, some values are assigned to the interior vertices (hierarchical levels) whereas in cladograms each edge may have some associated weight. Despite these obvious differences, it is useful to handle these OUC types together.

Matrix comparisons. The classical methods reduce the problem of evaluating dendrograms to the comparison of matrices. The idea is that each dendrogram may be replaced by a *descriptor matrix* **C** in which c_{jk} reflects the mutual relationship of objects j and k in the tree. This relationship, however, may be characterized in several ways as illustrated by Figure 9.6; and the choice among these descriptors is not always trivial. Podani & Dickinson (1984) listed the first five descriptors that follow; there is a sixth one, and it is possible that some other descriptors will also be introduced in the future.

1. *Cophenetic difference*: it is the lowest hierarchical level at which objects j and k belong to the same cluster (Fig. 9.6a). The levels pertaining to all possible pairs of objects are written into the cophenetic matrix **C** which is an unequivocal representation of the dendrogram (e.g., Sokal & Rohlf 1962). It means that the tree can be perfectly reproduced from **C**.

2. *Path difference:* the number of vertices along the path between objects j and k; it is one less than the number of edges connecting j and k (Fig. 9.6b, see e.g., Farris 1973, Phipps 1971, Williams & Clifford 1971). This descriptor has been referred to under various misleading names (topological difference, cladistic difference). Matrix **T** containing the pairwise path differences summarizes full information on tree structure (topology) but the hierarchical lev-

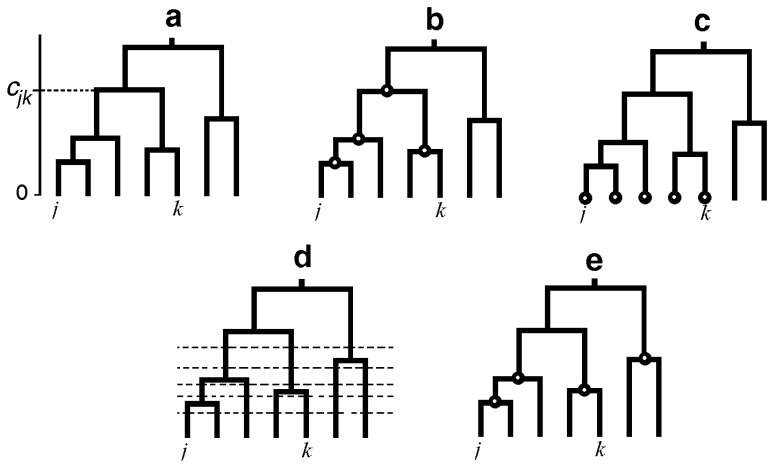

Figure 9.6. Dendrogram descriptors on the example of objects j and k in a dendrogram for 7 objects. **a:** cophenetic difference, **b:** path difference (= 4), **c:** cluster membership divergence (= 5), **d**: partition membership divergence (= 5), **e:** subtree membership divergence (= 4).

els are lost. It is a four-point metric, unsensitive to the position of the root and is therefore better suited to unrooted trees (Podani 2000).

3. *Cluster membership divergence* is the number of objects in the smallest cluster containing both j and k (Fig. 9.6c). The matrix **M** of divergences contains all information for the reproduction of tree topology (all values but the diagonal satisfy the conditions of being an ultrametric, m_{jk} is therefore a quasi-ultrametric, Podani 2000).

4. *Partition membership divergence:* This measure utilizes the property that a dendrogram is a series of nested partitions. Excluding the trivial case of all objects belonging to the same class, a dendrogram implies a maximum of m–1 partitions. This maximum is not reached if there are identical hierarchical levels or multifurcations. The relative position of objects j and k in the tree may be expressed by the number of partitions in which these two objects belong to different clusters (Fig. 9.6d). Although partition membership divergence is topological, the information concerning the sequence of hierarchical levels in the tree is also preserved in the m\m matrix of divergences. Therefore, this descriptor is best suited to ranked trees in which the absolute levels are replaced by their ranks.

5. *Subtree membership divergence:* This descriptor characterizes the tree based on its internal branching structure. In a binary tree (in which only dichotomies appear), there are m–1 subtrees, including the dendrogram itself. In fact, each interior vertex has its own subtree. The relationship between object pair j, k is measured by the number of such subtrees in which they do not occur together (Fig. 9.6e).

6. *Path length (patristic distance):* If the tree-generating procedure assigns a length (or weight) to each branch in the tree, then the sum of the lengths along the path between two objects provides a new descriptor, summarized in the *path length* matrix **P**. For the comparison of phylogenetic trees, path length is the most appropriate, although – if we forget about branch lengths – topological descriptors 2-3 and 5 may also be appropriate.

> Some of the descriptors are not new: cophenetic levels were discussed already when cophenetic correlation was introduced (Subsection 5.5.1) while patristic distances were described in the context of additive trees (Subsection 5.4.4). The six descriptors emphasize different properties of the tree, and they are therefore sensitive to different within-tree 'anomalies'. This must be kept in mind when selecting a particular descriptor for dendrogram evaluations. For example, if reversals occur in the trees, then cophenetic difference and partition membership divergence become meaningless. The presence of multifurcations has detrimental effects on the behavior of path length and subtree membership divergence. On the other hand, cluster membership divergence is not affected by reversals. When we wish to use cophenetic difference, the pattern of increases in the hierarchical levels should also be considered carefully. The dendrograms of Figures 5.7a and 5.11a, for example, differ considerably in this regard; the levels increase slowly in the first and 'exponentially' in the second. The use of cophenetic differences is not recommended in this case, because the two dendrograms are apparently not commensurable by levels.

Dendrograms and cladograms are compared by calculating the correlation or distance between their respective descriptor matrices. This univariate comparison is generalized to several descriptors as follows. Assume that the two dendrograms to be compared are denoted by D_1 and D_2. Their squared Euclidean distance based on five descriptors will have the following form:

$$\delta_{12}^2 = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \sum_{a=1}^{5} [\, x(a)_{1ij} - x(a)_{2ij} \,]^2 , \qquad (9.11)$$

in which the states of $x(a)$ represent the descriptors. For cladograms, descriptors 2, 3, 5 and 6 may appear in the third summation. Normalization is necessary to eliminate the inevitable scale differences among descriptors. Cophenetic levels are rescaled to fall into the interval [0,1] for each dendrogram. Cluster membership divergence is divided by m. Partition membership divergence may be normalized in similar way: the scores are divided by the number of partititons implied by the given dendrogram (max m–1). Path difference and subtree membership divergence are normalized by the actual maximum found in each dendrogram.

The following example is based on an extensive survey by the author (Podani 1985), and illustrates dendrogram comparisons in a complex design. The objective is to detect the relative impact of sampling (quadrat size) and data type upon the results of a phytosociological classification. The percentage cover scores of species were recorded in 20 quadrats, each containing a nested series of eight different sizes (in the manner shown in Fig. 1.9). The raw data were used to derive further three data types: two were obtained by the Clymo function (Formula 2.16a, c = 3 and c = 15), and the third was the ultimate simplification into the presence/absence form. The four types of scores may be arranged into a data transformation series, with two transitional stages between the quantitative and presence/absence types. The simultaneous change of data type and quadrat size provided 32 combinations, each characterized by its own 20×20 distance matrix and the dendrogram obtained from this by sum of squares agglomeration. The 32 distance matrices were then compared in all pairs using the correlation coefficient to yield a 32×32 matrix between these results. In its PCoA ordination (Fig 9.7a), axis 1 is very strongly unipolar and therefore uninformative. Axes 2 and 3, although explaining only 9 and 2.3% of the variation, respectively, are more interesting to us. The scatter diagram shows clearly that data type is more influential than quadrat size which is most negligible in the presence/absence case. The trends are less clear-cut, but still recogniz-

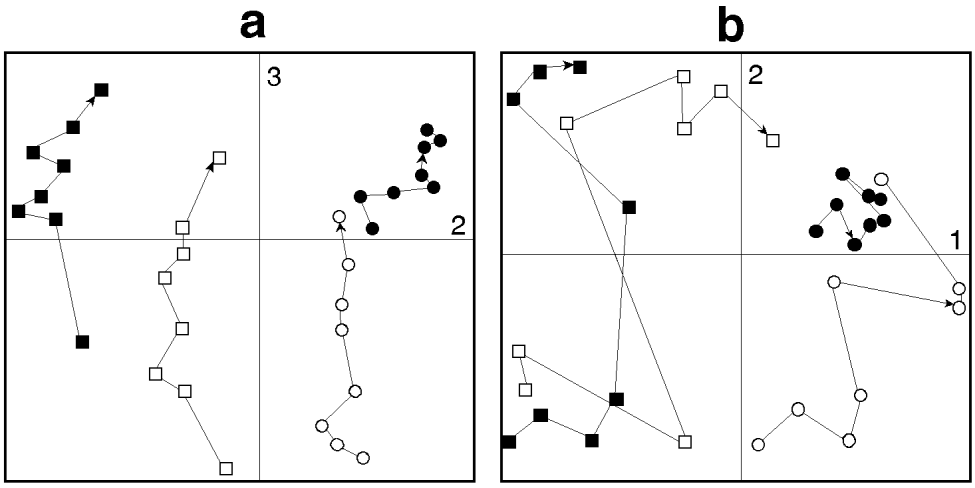

Figure 9.7. Complex comparisons to illustrate the joint effect of quadrat size (increases shown by arrows, from 0.25 to 16 m^2) and data type (■=cover, ❑= Clymo c=3, ○=Clymo c=15, ●=presence/absences). The PCoA ordination of distance matrices (**a**) shows the trends more clearly than the ordination of dendrograms (**b**). Redrawn after Podani (1989d).

able in the PCoA ordination of dendrograms from their matrix calculated using Formula 9.11 (Fig. 9.7b; the two axes explaining 21 and 16%). Along the first axis of this ordination, the first two as well as the second two steps of the data transformation series cannot be distinguished whereas the effect of quadrat size is the smallest in the presence/absence case, as before.

Graphical comparison. Two dendrograms or cladograms may be contrasted graphically using their respective descriptor matrices, as described in subsection 9.2.1.

Ultrametrics. A completely different approach to dendrogram comparison is due to Dobson (1975). The method examines the ultrametric inequality for each object triplet and then enumerates the number of triplets for which the inequalities are not the same in the two dendrograms. In other words, triplet $\{i, j, k\}$ counts if $c_{ij}<c_{ik}=c_{jk}$ satisfies in D_1 but we have $c_{ik}<c_{ij}=c_{jk}$ or $c_{jk}<c_{ij}=c_{ik}$ in D_2. This number may be divided by the possible maximum, $\binom{m}{3}$, i.e., the number of triplets for m objects, to provide the *ultrametric dissimilarity* measure which has the range of [0,1].

Branches and branch lengths. Another group of methods operates by counting the branches (edges) or by adding the associated lengths to derive tree dissimilarity measures. The basic idea is due to Robinson & Foulds (1979, 1981). The original approach was developed for unrooted trees although, with some modifications, they apply to dendrograms as well. The simplest index involves the *removal* of one branch of the tree at a time. In a dichotomous unrooted tree, the number of interior branches is m–3. The removal of either of them provides a two-cluster partition of the objects[1]. An interior branch of D_1 matches an interior branch in D_2 if their removal provides identical partitions. (For dendrograms, the two branches coming from the root must be treated as a single branch to allow the comparison.) The number of mismatching branches is then used as a measure of agreement between the trees (*symmetric-difference distance* or *partition metric*, Robinson & Foulds 1979, 1981). This number, divided by the possible maximum yields the edge matching coefficient for m>3.:

$$EM_{12} = \frac{\text{number of mismatched branches in } D_1 \text{ and } D_2}{2m-6}. \qquad (9.12)$$

Its range is [0,1]; 0 indicating full agreement, 1 corresponding to maximum disagreement. Measure 9.12 does not make any distinction between branches; their position in the tree or their lengths do not matter. However, if the sum of lengths of mismatched branches is divided by the total length of the two trees, then we have a weighted measure.

For unrooted trees, the best known measure is the *nearest neighbor interchange (nni) metric* (Waterman & Smith 1978) or *crossover* (Robinson 1971). We have seen in the discussion of cladograms that the interchange of subtrees pertaining to an interior branch is part of the optimization algorithms. According to the *nni* metric, the distance between two trees is the minimum number of such changes necessary to convert one tree into the other. For large m, determining the maximum is a formidable task, but Brown & Day (1984) described a fast approximation to the *nni* metric.

1 Branches leading to the terminal noda are disregarded, since their removal yields trivial partitions (one object plus all others) which always appear in both dendrograms.

9.2.4 Comparison of ordinations

Any ordination may be represented by the *t*-dimensional coordinates of *m* objects. In most comparisons, the first *p* (<<*t*) dimensions are of interest only because the others do not convey meaningful information (in sense of percentage variance, for example). Pairwise comparisons of ordinations, like any other OUCs, may be useful to assess the relative influence of variables, data types, resemblance functions and ordination algorithms upon the resulting configurations. The taxonomic congruence of ordinations of OTUs based on vegetative and reproductive characters may also be evaluated in this way. Since the description of shape in terms of coordinates is also an ordination (Subsection 7.6.2), the comparison of shapes by superposition methods is a special case for this approach. In one dimension, ordinations may be compared by the product moment or the rank correlation coefficient. In most cases, interest lies in at least two dimensions and correlations between coordinates do not work. Matrix correlation, however, may be a solution for two or more dimensions in such a way that each ordination is described by the distances of *m* points in the *p*-dimensional subspace (e.g., Podani 1989d, Figure 9.8). If such an approach is plausible, then the graphical comparison of two ordinations will also be possible. Notwithstanding the applicability of matrix correlation, a more elegant geometric procedure has received general acceptance in numerical ecology and morphometrics. This is the so-called *Procrustes method*, developed partly independently by several authors (Green 1952, Gower 1971a, Schönemann & Carroll 1970). The name refers to the ill-famed figure of Greek mythology, Procrustes the giant, who seized travelers in Attica and tied them to an iron bedstead by cutting off their legs or stretching them until they fitted it. Hence the expression, Procrustean bed which means "being forced to strict conformity under violent measures". The name reflects that the ordinations must be subjected to some drastic manipulations before any meaningful comparisons can be made. The basic assumptions of Procrustean analysis are that two ordinations are deemed indistinguishable if:

– either is obtained by shifting the other (i.e., by adding a constant to all of its coordinates);

– either of them is obtained via multiplying the coordinates of the other by a constant value;

– the rotation of either ordination by an angle α reproduces the other, including the special case of $\alpha = 180^{o}$ (reflection).

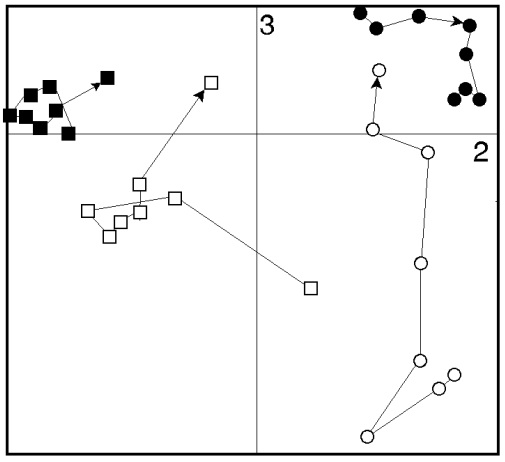

Figure 9.8. Meta analysis of ordinations using the same data as in Figure 9.7. A point represents a PCoA ordination of 20 quadrats for a given combination of data type and quadrat size. The PCoA of ordinations started from the complement of the matrix correlations for the first two original dimensions. Axes 2 and 3 are shown, with their relative percentages being 14 and 7%. Contrary to ordination 9.7a, the two extreme data types appear less sensitive to quadrat size changes (redrawn after Podani 1989d).

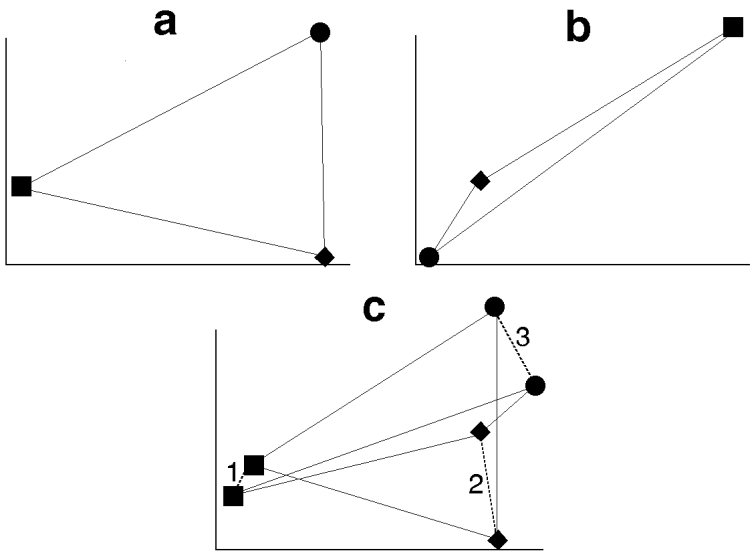

Figure 9.9. Procrustes analysis. The dissimilarity of two ordinations of three objects (**a** and **b**) is measured by the sum of squared distances (1+2+3) between the corresponding points in the best fit (**c**).

Starting from these assumptions, the comparison of two ordinations involves maximizing the fit of one ordination over the other by centring, rotation and dilation (rescaling). The best fit is then measured by the sum of squared distances between the corresponding points (Fig. 9.9).

More formally, if the centred coordinates of m points in p dimensions are written into matrices **X** and **Y**, then the minimum value of the following function is sought:

$$\sum_{i=1}^{m}\sum_{j=1}^{p}(x_{ij}-y_{ij})^2 = \mathrm{tr}\left[(\mathbf{X}-\mathbf{Y})'(\mathbf{X}-\mathbf{Y})\right] \qquad (9.13)$$

This is calculated by leaving **X** unchanged and transforming **Y** with the $p \times p$ orthogonal rotating matrix **H**:

$$\mathbf{H} = \mathbf{V}\mathbf{U}' \qquad (9.14)$$

in which **U** and **V** are derived from the singular value decomposition of **X'Y** (Appendix C):

$$\mathbf{X}'\mathbf{Y} = \mathbf{U}\mathbf{S}\mathbf{V}' \qquad (9.15)$$

(**S** is a diagonal matrix containing the square roots of eigenvalues). The goodness of fit is then expressed by the following formula:

$$R_E^2 = \mathrm{tr}\left[(\mathbf{X}-\mathbf{Y}\mathbf{H})'(\mathbf{X}-\mathbf{Y}\mathbf{H})\right] =$$
$$= \mathrm{tr}(\mathbf{X}\mathbf{X}') + \mathrm{tr}(\mathbf{Y}\mathbf{Y}') - 2\mathrm{tr}(\mathbf{Y}\mathbf{X}'\mathbf{X}\mathbf{Y}')^{1/2}. \qquad (9.16)$$

This is a symmetric measure, influenced heavily by the actual magnitude of coordinates. This undesirable property is eliminated by considering a multiplying factor c when matrix **Y** is rotated into $c\mathbf{Y}\mathbf{H}$:

$$c = \mathrm{tr}(\mathbf{Y}\mathbf{H}\mathbf{X}') \,/\, \mathrm{tr}(\mathbf{Y}\mathbf{Y}'), \qquad (9.17)$$

which leads to the following statistic:

$$R_S^2 = \mathrm{tr}(\mathbf{X}\mathbf{X}') - 2(\mathrm{tr}(\mathbf{Y}\mathbf{X}'\mathbf{X}\mathbf{Y}')^{1/2})^2 \,/\, \mathrm{tr}(\mathbf{Y}\mathbf{Y}'). \qquad (9.18)$$

However, this is no longer symmetric, therefore Gower (1971a) proposed to normalize the input ordinations to unit sum of squares right after centring:

$$\mathrm{tr}(\mathbf{XX}') = \mathrm{tr}(\mathbf{YY}') = 1. \tag{9.19}$$

It means that the squared distances of points from the origin in each ordination produce a unit sum. Formula 9.18 is then calculated from the normalized configurations; and the result is now abbreviated as d^2. As Sibson (1978) pointed out, R_S^2 may be normalized directly:

$$\gamma_S = R_S^2 \,/\, \mathrm{tr}(\mathbf{XX}'). \tag{9.20}$$

This measure ranges from 0 to 1. The relationship between d^2 and γ_S is:

$$d^2 = 2(1-\sqrt{1-\gamma_S}). \tag{9.21}$$

It follows immediately that the value of d^2 falls into the interval [0,2], 0 indicating perfect fit, 2 reflecting the maximum possible departure of one ordination from the other.

As an example. let us compare the PCA and COA ordinations of the objects (columns) of Table A1. According to the first two axes (Figures 7.2 and 7.14), the value of d^2 is 0.1, which seems quite low. We cannot make more statements on the result, however, until a reference basis is available for this comparison (see next section). It is noted that for the first three dimensions there is an inevitable increase of squared distances ($d^2 = 0.309$), the relative change indicating that the two ordinations differ most markedly along axis 3.

9.2.5 Comparison of rearranged matrices

The need of comparing rearranged matrices rarely appears in the literature, although the comparison of rearrangements obtained manually or via objective methods is as interesting as the evaluation of other OUC types. Here, the method developed for the comparison of cross-partitioned block classifications (Podani & Feoli 1991) is introduced briefly. The transformation metric between partitions (Formula 9.7) is modified for this purpose. Let $\mathbf{X}_i$ and $\mathbf{X}_j$ be two rearranged data matrices of size $n \backslash m$, with p clusters for rows and q clusters for columns in both. The first task is to determine the minimum number of rows and columns to be relocated in $\mathbf{X}_i$ to get $\mathbf{X}_j$ (or vice versa). The two row- and the two column-classifications are compared separately to derive the values of $M_{ij(rows)}$ and $M_{ij(columns)}$ (for brevity, *MINDMT* is replaced here by *M*). These provide the number of data values to be moved:

$$K_{ij} = mM_{ij(sorok)} + nM_{ij(oszlopok)} - M_{ij(sorok)}M_{ij(oszlopok)} \tag{9.22}$$

which may be divided by the possible maximum to obtain the κ index:

$$\kappa_{ij} = \frac{mM_{ij(rows)} + nM_{ij(columns)} - M_{ij(rows)}M_{ij(columns)}}{m\max M_{ij(rows)} + n\max M_{ij(columns)} - \max M_{ij(rows)}\max M_{ij(columns)}}. \tag{9.23}$$

Its range is [0,1], 0 indicating perfect agreement, 1 showing maximum discrepancy.

The comparison of matrices rearranged by *seriation* is achieved through the comparison of row and column permutations for data matrices, and row permutations for resemblance matrices. The strategy is that row-wise and column-wise dispositions are counted and then summed for data matrices, whereas only the rows are considered for resemblance matrices. Let the row and column indices in the first matrix be given by *i* and *j*, respectively, and the indices of the corresponding rows and columns in the second matrix be denoted by *y(i)* and *y(j)*. Then, the desired quantity will be obtained as

$$\kappa_{1,2} = \sum_{i=1}^{n} |i - y(i)| + \sum_{j=1}^{m} |j - y(j)|. \qquad (9.24)$$

Only the first term is used for resemblance matrices. A measure more elegant than this is analogous to *MINDMT* or the *nni* metric and is defined as the minimum number of *neighboring* rows and columns to be interchanged iteratively in the first matrix to obtain the second. As expected by the trained reader, the determination of this transitional measure is a much harder problem than calculating Formula 9.24.

9.3 Hypothesis testing, expectations and distributions

The pairwise comparison of results provides a dissimilarity measure which is either constrained to lie between fixed limits or has no theoretical upper bound (Function 9.11 is an example for the latter). The lack of upper bound poses no problems until we remain within the same meta-analysis such that the number of objects and other parameters of the survey do not change. Even the fixed upper bound, usually 1, is of little help whenever dissimilarity values coming from different surveys are to be contrasted. Can we surely say that a dissimilarity of $d^2 = 1.42$ between two ordinations of 40 objects implies greater disagreement than another value of $d^2 = 1.40$ calculated for two ordinations of 10 objects? Unexperienced analysts might say that d^2 ranges from 0 to 2 regardless the value of m, consequently the difference of 0.02 is a true indication of a slightly higher discrepancy between the 40-object ordinations. A statistically-minded investigator, however, cannot make such statements! Correct comparisons between dissimilarities can only be made if the reference distribution of the measure is known, together with all of its parameters, especially the expectation (mean). In the example above, for $m = 40$ the dissimilarity value of 1.42 is much below the mean, while for $m = 10$ the value of 1.40 is far beyond the expectation! Thus, a value of 1.42 implies a relatively strong agreement between 40-object ordinations, whereas 1.40 for 10-object ordinations indicates quite high dissimilarity. "Everything is relative", therefore any statement based on the numerical values only would be unwise. Furthermore, knowledge of the reference distribution is absolutely necessary if we wish to make a *significance test* of the dissimilarity measure. We may want to tell whether two OUCs obtained independently for the same objects are more similar than expected for random OUCs (*significant* result), or their dissimilarity falls into the range which characterizes most (usually 95%) of the randomly generated OUC pairs anyway (lack of significance). By addressing these questions, we reached a challenging and rapidly developing area of multivariate statistics.

Our problems are further complicated by the fact that the underlying distribution of most dissimilarity measures for OUCs is unknown. An exception is the partition metric (Formula 9.12) whose exact distribution has been derived up to 16 objects (Hendy et al. 1984). Some parameters of certain measures of cladogram dissimilarity are also known (Steel & Penny 1993). Usually, however, for practical problem sizes the distributions are not available or, if some theory is already developed, the computations are exceedingly difficult and impractical. The solution is offered by Monte Carlo simulation algorithms, or more precisely, a subset of these methods: the randomization and permutation tests.

The principal issue in Monte Carlo simulation is the formulation of a baseline situation corresponding to the null-hypothesis. When we sit down and think over the actual problem, it may turn out very quickly that the choice among different variants of Monte Carlo simulation is not as easy as earlier thought. Monte Carlo methods, in the strictest sense of the word, are used if the distribution should refer to randomly generated OUC pairs and we can say that any OUC is equally likely to occur in the random sample. Using an appropriate random number generator we simulate, say, 999 pairs of dendrograms with random hierarchical levels and entirely random bifurcations (the exact algorithm is not essential at this point, but see Lapointe & Legendre 1991, Podani 2000). The dissimilarity is calculated for every pair of dendrograms and then the 999 values are arranged into categories to draw a frequency histogram of dissimilarities. The actual dissimilarity to be tested, d, is the 1000^{th} value. Using the 1000 instances of the dissimilarity measure we may estimate the probability that for random dendrograms we get a dissimilarity less than or equal to d. If this probability is lower than the previously specified significance level α (usually 0.05), then the two actual dendrograms can be considered significantly similar and the null-hypothesis is rejected. For a random sample of 1000 and $\alpha = 0.05$, this happens if at least 950 of the simulated dissimilarities exceed d. In the opposite case, we retain the null-hypothesis by saying that the actual d value could be obtained for random pairs of OUCs (at the given α) and the similarity of the two dendrograms is not significant. The *permutation tests*[2] are based on similar grounds, with the only substantial difference being in the manner the sample OUCs are generated. In this case, the OUCs are not entirely random; only the objects are permuted by random relabeling. In other words, the arrangement of the objects is changed while the basic structure of the OUCs (a configuration of points in an ordination, a tree graph, etc.) remains constant. The so-called *exact permutation tests* derive the sampling distribution of the measure by generating all the possible permutations of objects, a strategy restricted to relatively small problem sizes. In practice, only an estimation can be made based on a limited number of random permutations. The larger this number, the better the aproximation to the 'true' distribution. For dendrograms, Lapointe & Legendre (1992) found that 1000 pairs provided a reasonably good estimate, while for matrices Jackson & Somers (1989) suggested as a rule of thumb a minimum of ten-to-hundred thousand simulations. In examining the distribution of measure 9.7, Podani (1986) found that 5000 pairs approximated very well the exact distribution. Clearly, there are no generally valid guidelines and much depends on problem sizes, but a process in which sample size is increased gradually along with repeated tests may be helpful to reach a stable result. Needless to say that permutation tests are typical computer-intensive procedures of contemporary statistical analysis.

The basic strategies of Monte Carlo and permutation tests are summarized for the major types of OUCs in Table 9.1. A "pure" Monte Carlo simulation is usually more difficult to achieve than simple permutation tests. The number of possible OUCs , from which the simulations derive a sample, is usually very high, often infinite, in sharp contrast to the relatively

2 The terms permutation test and randomization test are practically synonyms (Manly 1991). In the present case, the word permutation appears more straightforward, better indicating how the distributions are generated.

Table 9.1. Comparison of the strategies of Monte Carlo simulation and permutation tests for different types of results.

Result (OUC)	"pure" Monte Carlo simulation	Permutation test
Resemblance matrix	Entirely random resemblance values	Columns (and rows) randomly interchanged
Hard partition	Random assignment of objecs into k classes, regardless of their size	Random assignment of objects into k classes such that original group sizes are maintained
Fuzzy partition	Random weights for each object such that their sum is 1	The original weights are retained, the objects are randomly assigned to them
Dendrogram	Random levels, random bifurcations, randomly selected objects	The objects in terminal positions are randomly mixed
Rooted cladogram	Random branch lengths, bifurcations and object assignments	As above
Ordination	Random coordinates in every dimension	Original positions retained, objects randomly relabeled.

small number of possibilities in the permutation tests. The Monte Carlo methods are more general, while the permutation tests are suited to the actual circumstances.

9.3.1 Matrices

The difference between Monte Carlo simulation and the permutation-based strategy is illustrated through the significance test of matrix correlation, a method almost exclusively used for evaluating resemblance matrices **D** and **E** (Mantel 1967). Clearly, comparing the actual correlation with a threshold value found in a standard statistical table would be unwise because the values within each matrix are strongly interdependent. Instead, the rows (and therefore the columns) of matrix **D** are randomly permuted many times, and each 'perturbed' matrix is also correlated with matrix **E**. Then, the significance of r_{DE} is evaluated using the empirical distribution of the correlations coming from the permutations (Mantel test).[3] According to the null-hypothesis, the mechanisms generating the values in **D** are independent from those responsible for the structure implied by **E** so that it is likely to get r_{DE} even though one of the matrices is completely 'confused'. If it is not true, then the background mechanisms for the two matrices are not independent, the permutations destroy the basic structure and therefore the null hypothesis is rejected.

In addition to permutations, a test may be based on entirely random distance matrices. The keystone in this approach is to simulate distances guaranteeing that they correspond to some distance/dissimilarity structure, i.e., they are not mere random numbers. A possibility is to randomly permute the rows and the columns of the original data matrix (if available) and to calculate the correlation between the **D** matrices obtained from randomized data and the other

3 More precisely, it was the cross-products, rather than the correlations, that Mantel used in the computations. It is a reasonable choice because the cross-products are themselves sensitive to permutations, while the other terms in the correlation measure are invariant. If the entries in both matrices are standardized by standard deviation beforehand then the cross products will be equal to the regression coefficient. of **D** with repect to **E** and vice versa.

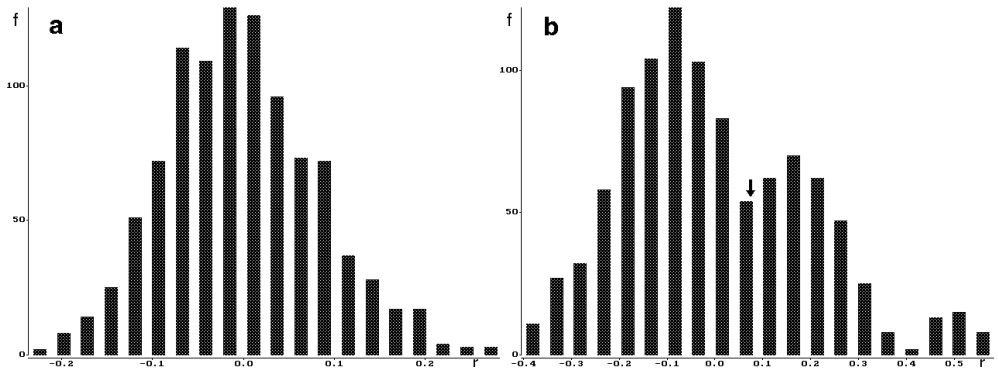

Figure 9.10. Frequency histogram of matrix correlation in a Mantel test based on 1000 permutations for **a:** distances calculated from the dune vegetation data for species and for environmental variables (Table A4); **b:** distance matrices calculated separately for monocots and dicots of Table A1. In the latter case, arrow points to the position of the test statistic in the distribution.

matrix **E** which remains intact, and to repeat the procedure many times. The distribution of the original dissimilarity measure is also of some concern, because it has some influence on the permutations (Hajdu 1981, Gower & Legendre 1986). In fact, the specific behavior of the formula should also be built into the simulation model, although the procedure may become too cumbersome this way.

As an example, let us examine the dune vegetation data again (Table A4). The distances among stands are calculated using species scores to provide the first matrix, whereas the second distance matrix is derived from the 'environmental' data. According to the null hypothesis, the two groups of variables are independent, so that the value of matrix correlation is less than or equal to 100(1-α)% of correlations obtained from permuted matrices. The actual value of matrix correlation is 0.44, which is higher than all the simulated values (Fig. 9.10a). Thus, the null hypothesis is rejected: the two groups of variables lead to significantly similar distance matrices, indicating dependence of species performance on the environment.

The second example is more artificial, yet useful to illustrate the opposite situation. Starting from Table A1, we examine whether the distance matrix of stands calculated for monocots (7 species) significantly correlates with another matrix based on the dicots (5 species). The correlation is 0.091, suggesting immediately that the two matrices have little to do with each other. This is confirmed by the permutation test: 31% of the simulated values are larger, 69% are smaller than 0.091 (Fig. 9.10b). In other words, every third permuted value is higher than the actual statistic so the null-hypothesis is accepted: the two groups of angiosperms provide matrices as dissimilar as the randomizations. This statement is valid for this example only; bear in mind that the validity of the Mantel test is restricted to the two matrices being compared!

9.3.2 Hard partitions

The distribution of measures of partition agreement is also best-examined by Monte Carlo simulations, even though in some circumstances some parameters could be derived exactly (cf. Hubert & Arabie 1985). Suppose that a d dissimilarity value for partitions P and Q is to be tested for significance. The number of clusters is s and t respectively, with cluster sizes p_i and q_j. In the plain Monte Carlo case, both the numbers of clusters and the cluster sizes are results

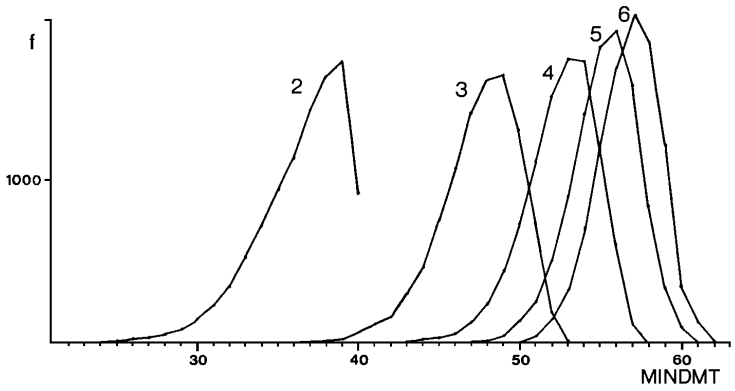

Figure 9.11. The frequency histogram of the partition agreement measure *MINDMT* obtained by Monte Carlo simulation for $m = 80$, $s = t = 2, 3, 4, 5, 6$. The points are connected only for clarity. Each distribution is based on the comparison of 10000 pairs of partitions (Podani 1986).

of random effects. However, it is more reasonable to keep at least s and t constant during the simulations. Even more attractive is the permutation test in which the random partitions are generated without changing the marginal values of the cross-classification table, i.e., p_i and q_j (Hubert & Arabie 1985).

The intermediate situation, with s and t fixed (and $s = t$ for all simulated partitions), is illustrated for the *MINDMT* measure for 80 objects (Figure 9.11). As seen, when s increases the expected dissimilarity also increases, whereas the distribution becomes less skewed. The effect of fixing cluster sizes is shown by the permutation based-simulations (Fig. 9.12). For comparison, the first example shows the most general case with free class sizes (Fig. 9.12a) for which the distribution agrees well with the histogram of Fig. 9.11, case of $s = 2$. At fixed class sizes, with gradually increasing the difference between the size of the two classes, the expectation decreases and the distribution becomes more symmetric (Fig. 9.12b-e).

It is worth examining the significance of changes depicted by Fig. 9.4. The number of species for which we have a 'significant' departure from the reference classification may be of interest. However, no formal statistical test can be made in this case because the partitions to be compared are not independent, being partly based on the same subset of variables. Nevertheless, the 'critical values' do provide a good basis for the comparison of functions of partition agreement. Let us choose the 95% 'probability level'. For each coefficient, the simulations provide the threshold value below which the two partitions are significantly similar. Cluster sizes were not fixed, because cluster sizes changed during species reduction. The threshold values are: 0.5 (MISC), 0.655 (1–LAS), 0.445 (1–RAND), 0.585 (1–JAC) and 0.415 (1–OCH). Now it becomes clear why are so diverging the values for the very same comparison. The *distribution* varies with the coefficient so that each value may only be compared to its own significance threshold, rather than to the values provided by other formulae. A fast scrutiny of the diagrams shows that species reduction to as low as 5 is not enough to modify the starting partition to an extent necessary for a 'significant'change for any coefficient.

9.3.3 Fuzzy partitions

Fuzzy partitions are described by cluster membership weights (coefficients of belonging) which provide a total of 1.0 for each object. In the most general Monte Carlo situation, entirely random weights could be generated such that this condition is met. The permutation tests, on the other hand, maintain the original values while permuting the objects. In other words, the rows of the weight matrix $\mathbf{U}_1$ (Section 4.3) are randomly mingled while the other matrix, $\mathbf{U}_2$, remains unchanged. The columns and therefore the column totals (the sum of

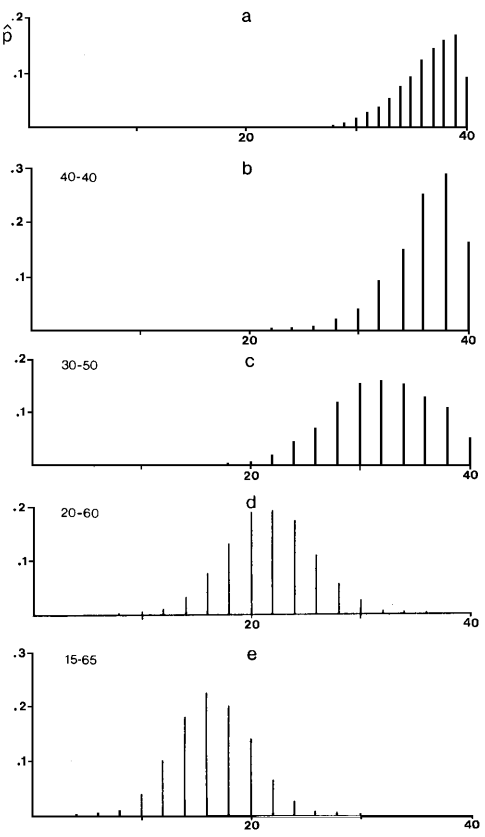

Figure 9.12. Estimated probability distribution of *MINDMT* for $m = 80$, $s = t = 2$ without fixing cluster sizes (**a**) and with fixed cluster sizes shown in the upper corner (**b-e**). Each histogram is based on the comparison of 10000 pairs of partitions (Podani 1986).

weights for each class) are not changed either to ensure compatibility with the permutation test of the agreement of hard (crisp) partitions. This topic is very little investigated and detailed simulation studies are in order.

9.3.4 Dendrograms and cladograms

Dendrograms and cladograms are perhaps the most complex mathematical objects among all types of OUCs encountered in the analysis of multivariate biological data. Their comparison, as we have seen above, as well as their simulation, as we shall see below, represent an intricate subject area. The topic is very far from being exhausted even in mathematics, but we know that there are a few basic assumptions that must be fulfilled for a correct test to be made. These include:

- The set of simulated trees is in fact a sample from the universal set of all possible trees. We must guarantee that each possible tree has the same chance of being selected. The question of what trees are in fact possible is always context-dependent (see below).
- The process of the simulation should be compatible with the formula to be tested. For example, if the measure is insensitive to the hierarchical levels, such as coefficients derived from path difference and subtree membership divergence, then the number of

possible dendrograms – and cladograms – is obviously V_m (Formula 5.16), and each of the V_m trees may appear in the simulation (e.g., Shao & Rohlf 1983). If the measure does reflect the order of levels (partition membership divergence), then the reference distribution should rely upon H_m different dendrograms (Formula 5.17) – a value considerably larger than V_m. The latter case is discussed by Lapointe & Legendre ("*double permutation algorithm*", 1991)[4] and – with less emphasis – by Steel & Penny ("D_{tip}" measure, 1993). The most difficult case is the simulation of hierarchical levels to provide a basis for testing cophenetic comparisons. Here, the number of possible dendrogram structures is again H_m, but randomly generated levels give rise to an infinite number of possible trees. Although Lapointe & Legendre (1991: 189) consider the possibility of such simulation, they admit that the best strategy is to restrict the possible hierarchical levels to those actually observed in the two dendrograms being compared.

- The above two paragraphs concern the plain Monte Carlo simulations. Their relationship to permutation tests is somewhat unclear at the moment. The comparison of two dendrograms follows the logic of Mantel tests only if the topologies and levels are fixed while the objects are permuted over the terminal vertices of either (or both?) of the dendrograms. Apparently, Lapointe & Legendre (1995) prefer the full randomization strategy against permutations. However, it is easily conceivable that we do not regard all possibilities to be equally likely in the random sample. Chained dendrogram shape is a case in point. Chaining is commonly observed with single link clustering while exceptional by complete linkage (recall the examples discussed in Chapter 5). Therefore, if single linkage is the strategy used, then balanced dendrogram shapes should be considered much less likely than chains and complete randomization would be "biased" in some sense.

Let us see a more concrete example. The dissimilarity between the two unrooted cladograms of Figure 6.1 may only be tested using the permutation-based approach. One tree is fully dichotomous whereas the other has many multifurcations, hence full randomization is not justified (Penny et al. 1993). Expressing tree topologies in terms of two path difference (PD) matrices, their distance is 126. After ten million permutations, this distance proved to be highly significant, because this value or an even lower distance occurred less than 100 times. In other words, the actual statistic is significant at the probability level of $p<0.00001$. The conclusion is that the language tree and the genetic tree are statistically similar – although the explanation of background effects is a different matter.

As a further illustration of permutation tests, group average clustering was applied to the two distance matrices used in the second example in Subsection 9.3.1. The resulting dendrograms are presented in Figure 9.13. Since the distance matrices were not significantly similar, we can expect that the dendrograms derived from them will not be similar either. We calculate two values, the matrix correlation between the respective PD matrices and between the partition membership divergence (PMD) matrices. The results are –0.01 and –0.22, respectively. In the histogram based on 1000-1000 simulated values obtained for entirely random dendrograms, these statistics fall into the acceptance region at the probability level of $\alpha = 0.05$. Nevertheless, the two correlations do not imply the same thing which turns out if we carefully examine the distributions. The PD-based correlation is almost equal to the mean of

4 The dendrogram simulator routine of **SYN-TAX** uses the random agglomeration strategy which is computationally more efficient (Podani 2000).

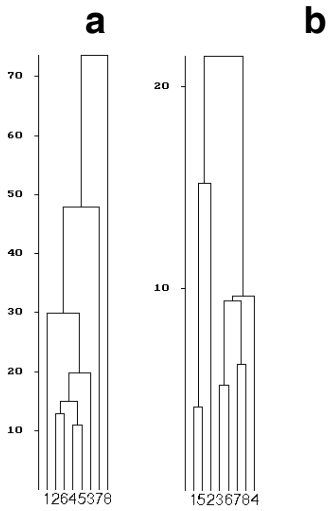

Figure 9.13. Group average clustering of sample sites (Table A1) from Euclidean distances calculated using monocots (**a**) and dicots (**b**).

simulated values (–0.004), whereas the correlation using partition memberships is very close to the threshold of significant difference (at $\alpha = 0.05$, the simulated threshold is –0.267 while the mean is 0.004). In other words, regarding the branching pattern the two dendrograms differ to the extent expected for two random ones whereas according to the orders their large distance is a rare event even for random dendrograms We may conclude that measures of dendrogram distance do not say much by themselves; knowledge of the underlying distribution increases interpretability of the statistics.

9.3.5 Ordinations

Random ordinations are easier to generate than dendrograms. For example, Podani (1991) suggested to simulate random and uniform coordinates for the objects in the pre-selected k dimensions to serve as a basis of hypothesis testing. Multivariate normality of random coordinates or other Monte Carlo models are also conceivable in the derivation of a random sample of ordinations. In these cases, we do not have to worry about the scale on the axes, because Formula 9.21 of Procrustes analysis implies normalization to unit sum of squared distances from the origin.

Permutation tests maintain the original coordinates while relabeling the points randomly. Of course, this operation implies a null hypothesis completely different from the plain Monte Carlo case, because the *only* point scatter allowed in the simulations is the actual one. This must be remembered when evaluating the test statistics. However, as the following example demonstrates, the discrepancy between full randomization and permutation decreases when the number of dimensions is increased.

The PCA and COA ordinations of objects of Table A1 were already compared at the end of Subsection 9.2.4. Let us now generate the reference distribution of d^2 for two and three dimensions based on full randomization using the uniform distribution (Fig. 9.14a) and on random permutations (Fig. 9.14b). Beware that the two ordinations being compared are not independent since they are calculated from the same data. Thus, we do not perform a formal significance test as such; the simulated distributions will only be used to assess the departure

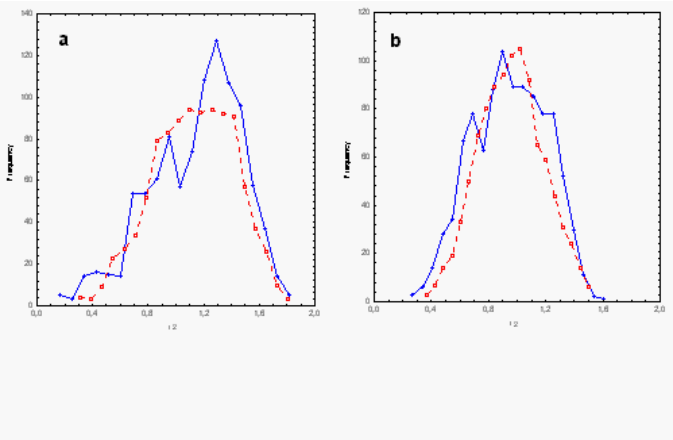

Figure 9.14. The distribution of Procrustean dissimilarity (d^2, Formula 9.21) between two ordinations of 8 objects in two (**a**) and three (**b**) dimensions. Dotted lines (fully random case) and solid lines (permutation) are used to improve clarity of the histograms.

of d^2 from the expectation, thus achieving some 'normalization' of the coefficient. The simulated means are 1.14 and 0.95, respectively, for two and three dimensions, regardless the simulation strategy. The comparison of actual dissimilarities (0.1 and 0.309) to these expectations shows that in two dimensions the two ordinations are relatively more similar than in three dimensions, much more than one might think by judging the absolute difference between the values (0.209). The major difference between the simulated histograms is that the ranges are slightly wider for the permutations (solid lines, Fig. 9.14) than for the random uniform strategy.

9.3.6 Unplanned comparisons – in general

The examples discussed thus far share a common property: when there were more than two OUCs compared simultaneously we did not perform any significance test of their dissimilarities; hypothesis testing was restricted to single pairs only. We had a good reason to do so: statistical analysis and the choice of the probability level in multiple comparisons must be done very carefully. Suppose that we have k results to be compared, so that the total number of pairs is $g = k(k–1)/2$. If we select certain pairs *a priori* such that they are independent (OUC1 vs OUC2, OUC3 vs OUC4, and so on), then the test may be based on the simulated coefficient as decribed previously. However, if we do not know in advance which independent pairs are of interest to us, but rather we wish to select 'significantly' similar OUC pairs from the set of g pairs (*a posteriori* test), then the usual thresholds do not apply because of the accumulation of Type I error. If the test were made in the usual manner, then more values would be regarded (erroneously) to be significant than there actually are at the given probability level. To avoid this, a more rigorous test is to be made by increasing the threshold dissimilarity (right-tailed tests) or decreasing it (left-tailed tests). There are several possibilities to accomplish this, but we show only two:

- The threshold is redefined such that the total Type I error for all the g comparisons does not exceed α. In order to do it correctly, the probability level for a single pair must be lowered according to

$$\alpha' = 1 - (1-\alpha)^{1/g} \qquad (9.25)$$

whereas the reference distribution is the same as for two OUCs. This formulation has been originally suggested for the multiple comparison of group means (cf. Sokal & Rohlf 1981a).

- Another possibility that deserves mention is the simulation of the distribution of extreme values. In the general Monte Carlo model, a set of *k* random OUC is generated while in the permutation case, all OUCs in question are permuted (randomly relabeled). Then, comparisons are made in all possible pairs and the extreme dissimilarity (usually the minimum) is found. This is repeated many, say, 1000 times to obtain the *distribution of extreme values* derived from comparisons that are not independent. After selecting a probability level α, we identify the associated threshold in the histogram of extremes. This is used in testing whether *any* value from the *g* comparisons of original OUCs is significant.

Let us examine the effect of the above restrictions upon the comparison of partitions using the *MINDMT* formula. Assume that there are 80 objects, divided into two clusters in each of five independently derived classifications. In the planned strategy with $\alpha = 0.05$, from the distribution on the left of Fig. 9.11 we have that the threshold is *MINDMT* = 31, so that an actual dissimilarity equal to or lower than this value would indicate significant similarity between a given pair of partitions. The five classifications form 10 pairs whose comparisons cannot rely upon the same critical value. From the simulated distribution of extremes (minima), at α = 0.05 we obtain that the adjusted threshold of *MINDMT* is 27. Consequently, we shall find fewer pairs of OUCs to be significantly similar than would otherwise be detected by disregarding the accumulation of Type I errors. Note that, for ten pairs, Formula 9.25 provides α' = 0.005, leading from the simulation of the distribution of *MINDMT* (Fig. 9.11) to the same critical value (27).

9.4 The consensus approach

The term *consensus*, quite familiar for us from everyday political declarations, refers in biology to a synthesis of *k* alternative and equally important results derived for the *same* set of objects. This new result emphasizes agreements among the competing OUCs and is usually considered to be a more adequate representation of inter-object relationships than any starting OUC by itself. This is partly because consensus generation may eliminate the effects of our subjective choices regarding the number of variables, data types, resemblance functions and clustering or ordination procedures necessarily made during processing of our data. Also, whenever a procedure yields several different, yet equally optimal final results (e.g., cladograms), the only resolution is their synthesis into a new result. A first glance at the vast literature of relevant methods suggests that although consensus generation is conceivable for any kind of results of multivariate data exploration, hierarchical classifications represented by dendrograms and cladograms have received the most attention.

9.4.1 Consensus partitions

First, hard partitions containing disjunct, non-overlapping classes will be considered. The consensus method aims to synthesize $k \geq 2$ partitions of *m* objects – in which the number of classes is not necessarily the same – into a new partition. Although there have been several at-

tempts to find a single consensus partition (see below), Neumann & Norton (1986) pointed out that the consensus problem usually has several, equally acceptable solutions. All of them can be derived from the so-called *strict* consensus defined as the unique partition in which any class j contains only those objects that belong to the same cluster in all of the k starting classifications. This is not contradictory with any initial partition but nevertheless has the potential disadvantage that there may be many, even m consensus groups if the differences among partitions are high. Therefore, the practical utility of strict consensus partitions is often questionable. Successive fusions of classes of the strict consensus provide a series of *intermediate* consensus partitions in each group of which the objects occurred together at least in k–1, k–2, k–3,... partitions. Ultimately, these fusions provide the other extreme synthetic classification, the *loose* consensus. In this, all objects that belong together in at least one of the initial partitions appear in the same group. The disadvantage of loose consensus is that the presence of a few objects of uncertain group membership leads to a single trivial consensus group. Two or more loose consensus clusters indicate that no member of any group was ever clustered together with any object from any other group, i.e., the consensus clusters are fully *isolated*.

To illustrate the above consensus series, let us consider the following sample partitions of 10 objects:

$$\begin{aligned} P_1 &= \{1, 2, 3, 4\}\ \ \{5, 6, 7, 8, 9, 10\} \\ P_2 &= \{1, 2, 3, 4, 5, 6\}\ \ \{7, 8, 9, 10\} \\ P_3 &= \{1, 2, 3, 4, 5\}\ \ \{6, 7, 8, 9, 10\} \\ P_4 &= \{1, 2, 3, 7\}\ \ \{4, 5, 6, 8, 9, 10\} \end{aligned} \qquad (9.26)$$

Their strict consensus partition is given by:

$P_s = \{1, 2, 3\}\ \{4\}\ \{5\}\ \{6\}\ \{7\}\ \{8, 9, 10\}$

There are several intermediate consensus results, for example:

$P_c = \{1, 2, 3\}\ \{4, 5, 6\}\ \{7, 8, 9, 10\}$,

but each starting partition could have also been mentioned as an intermediate OUC! Finally, one easily verifies that the loose consensus is a trivial one, because all objects are assigned to a single group.

The above example demonstrates convincingly that the number of possible consensus partitions can be too large even in relatively simple situations. Nevertheless, the consensus candidates do not appear equally meaningful. It is of particular interest to derive consensus partitions that reflect agreements between more than 50% of the alternatives. Such a *majority rule* does not work for two classes in the example, because object 5 has a very ambiguous position (and the application of the majority rule is less straightforward for small and even values of k). For three clusters, however, we can easily find the consensus partition:

$P_t = \{1, 2, 3, 4\}\ \{5, 6\}\ \{7, 8, 9, 10\}$

The objects of each cluster in P_t appear together in at least three competing partitions. This is at the same time an intermediate consensus and we cannot be sure that there is always a unique majority rule consensus. Furthermore, the 50% threshold is just one, although important rule, and one may wish to set the threshold to be any percentage larger than 50. Another possibility of selecting from the intermediate consensus paritions is to search for the *median* consensus

partition (cf. Barthélemy & Monjardet 1981). This requires definition of a d dissimilarity function measuring partition agreement. Then, a given partition P_m is the median consensus of the k partitions if the following condition is satisfied:

$$\sum_{i=1}^{k} d(P_m, P_i) = \min_c \sum_{i=1}^{k} d(P_c, P_i) . \qquad (9.27)$$

Index c refers to any intermediate consensus partition. In a sense, the median consensus is on the average the closest to the alternatives but again, we cannot be sure that there is a unique solution for criterion 9.27.

Determining the strict consensus is an easy task, while obtaining the majority rule and median results is a more difficult (in fact, *NP*-hard) problem, especially for large numbers of objects. As a practical heuristics, we can apply an agglomerative, hierarchical consensus generation procedure (Podani 1989a). The analysis starts from distance matrix $\mathbf{D}_{m,m}$ with d_{jk} being the number of partitions in which objects j and k do *not* belong to the same class. The global optimization strategy (Subsection 5.2.4) is a straightforward clustering procedure in this case because the strictness of the consensus (i.e., within-cluster average distances) and the isolation of objects (between-cluster average distances) are simultaneously measured. The hierarchy obtained is a series of intermediate consensus partitions from which the consensus for a particular number of groups is easy to determine. A conceptual advantage of the hierachical consensus is its ability to show that several consensus results may exist for a given set of k alternatives. Diday & Simon (1976) – without reference to consensus generation – have proposed earlier to subject matrix **D** to complete linkage clustering.

The hierarchical consensus of non-hierarchical classifications is illustrated by an actual example (Podani 1989a). A set of eighty vegetational sample plots from the dolomite rocks of Sashegy, Budapest) were classified into three classes by six different clustering procedures. The task is to find the consensus partition which, if superimposed on the map of the area, provides a more generally valid vegetation map than any starting classification (Fig. 9.15). Clusters A, B and C may be identified on the map as the open *Festuca*-dominated community, the *Bromus* grassland and the *Sesleria*-dominated closed grassland, respectively.

To *fuzzy partitions,* the median consensus easily applies (Podani 1990). The median consensus of k fuzzy partitions is defined as a fuzzy partition with a minimum sum of squared diffrences in membership weights from the others. Let u_{ijh} denote the group membership weight of object j for cluster h in partition i. Furthermore, let u_{cjh} be the weight in the consensus partition sought (Fc). If the number of classes is p in every partition, then the objective is to minimize the quantity

$$SSQ_c = \sum_{i=1}^{k} \sum_{j=1}^{m} \sum_{h=1}^{p} (u_{ijh} - u_{cjh})^2 . \qquad (9.28)$$

It is found by exhaustive search, that is, the k classifications are fitted to one another in all the possible permutations of clusters. Dummy classes are added when necessary to allow comparison of partitions with unequal numbers of classes. There are $p!^{k-1}$ different permutations, so that the search is not operational for many classes or many partitions. The centroid method has

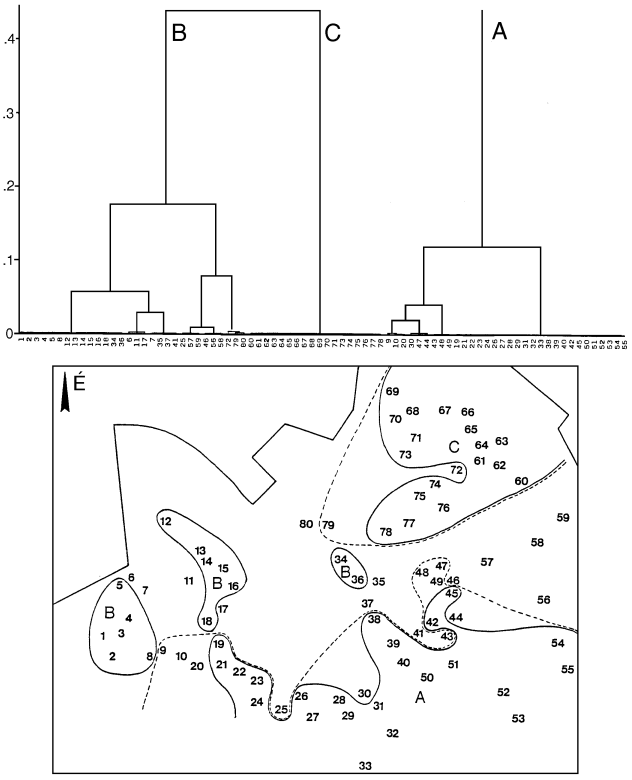

Figure 9.15. Hierarchical consensus of partitions: global optimization clustering from six 3-cluster partitions of 80 sample sites (top), and the projection of three major groups of the strict consensus (solid line) and the nearly optimal 3-class majority rule consensus partition (doted line) onto the map of the study area (bottom).

been proposed (Podani 1990) to find an approximation to the optimum when exhaustive search is impractical.

Since the hard partitions are special cases of fuzzy classifications, a fuzzy classification as a summary of hard partitions appears a very natural choice. Hard partitions, as their name suggests, are not flexible enough and cannot be modified to show slight details in the consensus object. Often, they require too many groups in order to be unique (recall the case of object 5 in the profile 9.26). On the other hand, a fuzzy synthesis of hard clusters may reflect minor details because the cluster membership weights are measured on a continuous scale.

The exhaustive search for a fuzzy consensus of partitions in profile 9.26 provides the following cluster membership weights to two classes (the matrix is transposed):

1.0 1.0 1.0 0.75 0.5 0.25 0.75 0.0 0.0 0.0
0.0 0.0 0.0 0.25 0.5 0.75 0.25 1.0 1.0 1.0

9.4.2 Consensus trees

The summarization of alternative results into a consensus is perhaps the most challenging task in contemporary molecular systematics. Regardless whether dendrograms or cladograms are used, only the topological structure of the tree is considered in most cases, whereas the hierarchical levels and branch lengths are disregarded (with noted examples, see below). Thus, it is almost always immaterial whether the OUCs are dendrograms or cladograms. The consensus

trees represent a compromise in a sense that the strong condition of allowing bifurcations only has to be released, except in trivial situations. This is best demonstrated by the *strict consensus trees* (Sokal & Rohlf 1981b, Swofford 1991) which – in agreement with the strict consensus partitions – are constrained to show those clusters only that appear in all competing trees. In other words, if a particular group of objects appears in the consensus tree, then they were always classified together in the input trees. Therefore, the interpretation of a strict consensus tree is fairly easy.

> Let us examine trees a, b and c of Fig. 9.16, summarized into the strict consensus tree d. The cluster {A, B, C} is the only one that occurs in all the three dendrograms, illustrating a disadvantage of most strict trees: the proliferation of politomies. In extreme cases, as for the cladograms of Fig. 6.18, all branches of the strict consensus tree originate directly from the root (= 'bush' or 'star tree'), which is not a very attractive property. In fact, the fewer the politomies in the consensus tree, the more similar are the input trees. This is measured by a consensus index discussed in 9.4.2.1.

The *semi-strict* or *combinational consensus* (Bremer 1990, Swofford 1991, Quicke 1993) implies some more relaxed conditions. The fundamental requirement here is that no clusters of the consensus tree should conflict with the starting trees. Since the presence of {A, B, C} in tree c (Fig. 9.16) does not contradict with group {A, B}, the consensus tree will be dichotomous for these three objects (Fig. 9.16e). If all input trees are fully dichotomic, then the strict and the semi-strict consensus trees will be identical. The *majority rule* consensus tree

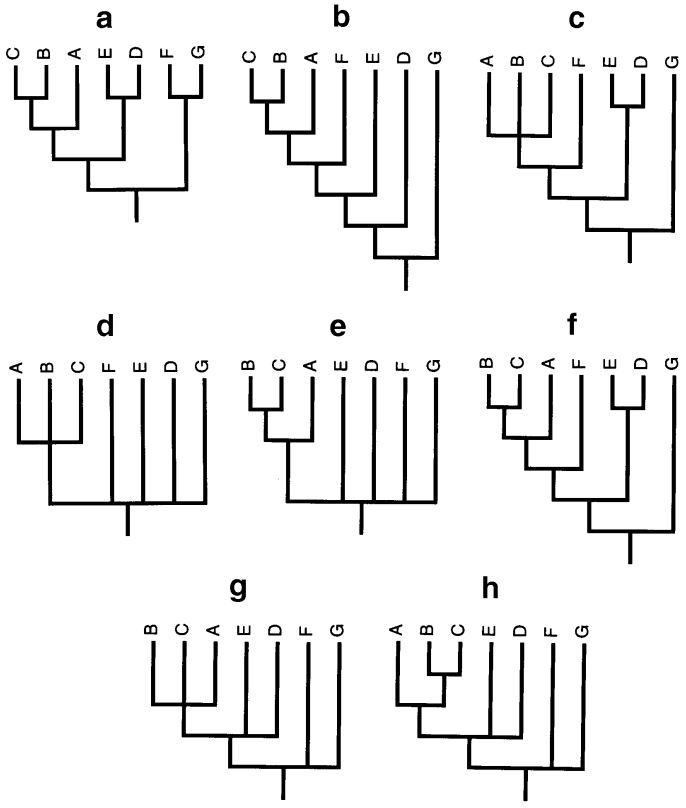

Figure 9.16. Consensus trees. The summary of alternative trees a-c into the strict consensus d, the semistrict consensus e, the majority rule consensus (>50%) (f), the Adams consensus (g) and the "durchschnitt" consensus (h). Compare the success of the consensus trees in finding the compromise tree for dendrograms a-c!

(Margush & McMorris 1981) tolerates even more discrepancies among the trees. The condition of the appearance of a consensus cluster is its presence in at least p per cent (p>50%) of the input trees. Consequently, the majority rule consensus for the above example will be completely bifurcating (Fig. 9.16f): its classes ({BC}, {ABC}, {ABCF}, {DE}, {ABCDEF} and {G}) are recognized in at least two of the three starting dendrograms. The value of p is selected by the investigator, and for many trees it is useful to raise the tolerance level well over 50%. It is easy to verify that for two trees the strict and the majority rule consensus trees are identical. As a further possibility, the *median* consensus tree (Barthélemy & Monjardet 1981, Barthélemy & McMorris 1986) also requires attention, Its derivation is based on the same grounds as the median consensus partition (9.27) provided that we find an appropriate function for the pairwise comparison of trees. If this function is the partition metric (Equation 9.12), then the 50% majority rule tree is at the same time a median tree (Barthélemy & McMorris 1986), which is not necessarily bifurcating. If one insists to find a completely dichotomous median tree, the suggestions by Penny et al. (1982) should be considered to get a '*median binary tree*' (Swofford 1991).

It is an interesting historical fact that the first proposition for a consensus tree differs from all of the above-mentioned, mathematically elegant procedures. Adams (1972) suggested to examine how the large groups are subdivided into smaller and smaller clusters when we proceed from the root towards the terminal branches of the tree. First, we generate the partitions determined by the first division (i.e., at the root) and find their strict consensus partition. For dendrograms 9.16a-c, these partitions are {ABCDE}{FG} and {ABCDEF} {G} twice. From these, we obtain the strict consensus partition {ABCDE}{F}{G}, so the consensus tree starts with a trifurcation. Afterwards, the cluster {ABCDE} is evaluated in the same way, with a result shown in Fig. 9.16g. The most common criticism against the Adams tree is that it may depict clusters that did not appear in any of the starting trees (Fig. 9.17).

There is a method that applies exclusively to dendrograms, the '*durchschnitt'* (cross-section) consensus (Neumann 1983, Smith & Phipps 1984). The procedure relies heavily upon the ordering of hierarchical levels. From the root towards the leaves, partitions are defined by 'cutting' the tree, and then these partitions are summarized into a consensus at each cut level. The clusters of this consensus partition form the branching pattern of the consensus tree. The procedure may continue this way until the trivial partition of objects into m classes is obtained (Fig. 9.16h). Again, the actual hierarchical levels do not matter, only their

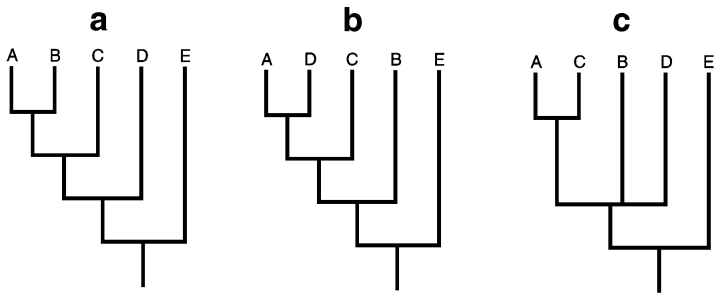

Figure 9.17. A major 'problem' with the Adams consensus tree (**c**) is that some of its clusters may be absent from all competing trees (**a-b**). Nevertheless, this tree expresses faithfully the relative neighbor relationships of most object pairs

ordering is considered. The durschnitt consensus can also have clusters that did not appear in the input trees at all.

Finally, I mention the pruning and grafting method proposed by Finden & Gordon (1985). Its major difference from all previous consensus tree generation methods is that some of the objects may not appear in the final tree. The method aims to remove outlier or conflicting branches in a stepwise manner such that the remanining truncated tree agrees with all input trees. Of course, several such reduced trees may exist for a given problem, and the one with the most numerous objects ('*largest common pruned tree*') is to be retained as the final consensus. The method is extremely useful if a few objects with highly unstable positions are responsible only for the between-tree differences. The problem is that we do not know exact and fast algorithms to identify the largest common pruned tree; enumeration of all possibilities for large m is not feasible for theoretical resons (cf. NP-completeness).

9.4.2.1. Consensus indices. When the overall agreement of many trees is measured by a single number, rather than a tree, we use a consensus index (Rohlf 1982). Indeed, these formulae express similarity between 0 and 1 and could have been treated for the case $k = 2$ in the subsection on pairwise dendrogram comparisons (9.2.3). However, these are most appropriate for many alterative trees. A good summary, in addition to Rohlf's review cited above, is in Swofford (1991) while the following discussion is confined to a few of them.

The simplest of all is the *consensus fork index* proposed by Colless (1980). It measures the deviation of the consensus tree from the fully binary one. The index is the number of non-trivial[5] classes divided by m-2, the maximum number of non-trivial classes. For dendrogram 9.16d, the index amounts to 0.2 (because only one class out of the five possible ones appears). Its value is 0.4 for tree 9.16g and 1 for tree 9.16f. The Mickevich-index (1980) assigns a weight to each consensus class, according to its size, thus representing an extension of the Colless-index. If cluster i has n_i objects, then its importance is $N_i = \min\{n_i–1, m–n_i\}$. The sum of these importance values divided by the possible maximum of the index provides the measure sought. For the three dendrograms mentioned above, we obtain the values of 0.222, 0.444 and 0.888, respectively. Finally, Schuh & Farris (1981) offer a completely different weighting system: compute the number of object pairs for each cluster, that is $N_i = n_i(n_i – 1)/2$, and add them ('*levels sum')*. For the example trees of Fig. 9.16 d, g and f we get 3, 26 and 13. The levels sum could be divided by the maximum to have a unit range, A problem with this ranging by the maximum for all indices is that the maximum depends greatly on the shape of the trees (larger for chained dendrograms than for balanced trees).

9.4.3 Consensus ordinations

If there are k alternative ordinations of the same m objects, then their average[6] (or, in a sense, their consensus) ordination may also be interesting for the investigator. The formulation of our task appears relatively easy because it seems sufficient to adapt the principle of median

5 A class is trivial if it contains only one or all objects.

6 An important field of application of consensus ordination is in morphometric analysis, already mentioned in Subsection 7.6.2 under the term 'superposition methods'.

consensus: the ordination sought is a point configuration whose sum of squared differences from all the others is the minimum (Formula 9.13). The objective is thus to optimize the fit of ordination k+1 to the k alternatives, achieved by the *generalized Procrustes method* suggested by Gower (1975). As many other multivariate optimization techniques, this method also requires several iterations to find the final solution. An obvious exception is the case $k = 2$ because the consensus coordinates are derived simply by averaging after the two ordinations are fitted, and this operation requires a single comparison (as described on page 329).

In the first step of generalized Procrustes analysis, each ordination is centred and normalized to unit sum of squares. Without doing so the concept of an average ordination would not work. If $\mathbf{X}_i$ denotes ordination i, and $\mathbf{Y}$ refers to the consensus configuration we are looking for, then each $\mathbf{X}_i$ configuration must be rotated to obtain its best fit to $\mathbf{Y}$, providing the result in $\mathbf{Y}_i$:

$$\mathbf{Y}_i = \rho_i \mathbf{X}_i \mathbf{H}_i \,. \tag{9.29}$$

In this formula, ρ_i is a scale parameter and $\mathbf{H}_i$ is the rotation matrix obtained by minimizing the following function:

$$RES = \sum_{i=1}^{k} \text{tr}\,[(\mathbf{Y} - \mathbf{Y}_i)'\,(\mathbf{Y} - \mathbf{Y}_i)] \tag{9.30}$$

(*RES* denotes the residual sum of squares). The rotations are performed such that the total sum of squares of the original k ordinations does not change:

$$SSQ = \sum_i \text{tr}\,(\mathbf{X}'_i \mathbf{X}_i) = \sum_i \text{tr}\,(\mathbf{Y}'_i \mathbf{Y}_i) = \sum_i \rho_i^2 \,\text{tr}\,(\mathbf{X}'_i \mathbf{X}_i)\,. \tag{9.31}$$

The consensus ordination is thus the arithmetic average of the respective coordinates:

$$\mathbf{Y} = 1 / k \sum_i \mathbf{Y}_i \tag{9.32}$$

So far so good, but how to fit each ordination to the consensus when the consensus is not yet known? To get out of the vicious circle, we have to iterate. First, $\mathbf{X}_2$ is fitted to $\mathbf{X}_1$ and then $\mathbf{X}_3$ is fitted to the average of $\mathbf{X}_2$ and $\mathbf{X}_1$. We proceed in similar way until ordination $\mathbf{X}_k$ is fitted to the average of all other ordinations. This first cycle yields the starting estimate of the consensus ordination. Usually, further cycles are necessary to improve this configuration; the iterations are stopped when the change of *RES* between two subsequent cycles becomes negligible. Each step involves rotations and an optional, though strongly recommended rescaling of coordinates. The directionality of the final consensus ordination is arbitrary. It is suggested therefore to perform a PCA from the final $\mathbf{Y}$, and then fit each starting ordination to this PCA result again.

The total sum of squares (*SSQ*) has two components: the residual sum of squares (*RES*) and the consensus sum of squares (*SSQ–RES*). The worse the overall fit of ordinations, the higher is the value of *RES*. A useful interpretational vehicle is the percentage contribution of each object and each ordination to the value of *RES*; the percentages identify outlier objects as well as ordinations that differ most remarkably from the average configuration.

As an example, let us see how the effect of plot size can be eliminated from an ordination of vegetation data. The study design was already mentioned in Subsection 9.4.1, but in this case we use only six sizes (from 1.5×1.5 m^2 to 4×4 m^2, with a side length increment of 0.5 m) for the same 80 sample sites. The ordination method is PCoA. The overall fit of the six PCoA results for the first two dimensions is obtained by the generalized Procrustes method. Since the scatter diagram would be too complicated and difficult to view if all the points were shown, the ordination outlines only the positions of a given site when quadrat size was changed. The consensus positions are not illustrated either; these are near the centroid if each shape. Figure 9.18 demonstrates that changing the quadrat size did not excert too much influence upon the ordination as a whole. The arched arrangement of quadrats along the back-

Figure 9.18. Generalized Procrustes analysis. The six ordinations to be compared were derived from different plot sizes for the same set of 80 sites. Each irregular shape represents the outline of positions for the same site. The individual points are not shown, so that the consensus ordination is only implicitly present in the diagram.

ground gradient from the open to the closed community type is essentially unaffected by quadrat size. This is now the right place to explain why the smallest two sizes (0.5×0.5 and 1×1 m^2) were removed from the analysis: they were too small to provide sufficient information on the species composition of the site and therefore the inclusion of their PCoA ordinations completely confounded the consensus ordination.

9.5 Comparison of results of the different type

9.5.1 Numerical comparisons

OUCs of the different type can only be compared numerically if all of them are brought into the same mathematical form. This universal standard is a *symmetric matrix*, summarizing the relationships of the m objects in all possible pairs. Two matrices can then be contrasted by the correlation coefficient (Formula 9.1), whereas Euclidean distance and related functions are useful only if the values of the two matrices are normalized to the same scale. Rank correlation is a possibility if we do not worry about the actual differences between the values in the matrices. An example for this approach was already presented in Subsection 5.5.1: the cophenetic correlation measures the distortion implied by the ultrametric tree in comparison to the original dissimilarities from which the dendrogram was obtained. Analogously, the correlation between the distances in an additive tree and the starting distances may also be calculated to measure the deviation of within-graph distances from the original distances. Since partitions and ordinations may also be written in form of $m \backslash m$ matrices, the correlation formula applies to a wide variety of combinations of OUC types.

9.5.2 Graphical comparisons

When all results are expressed in matrix form, the coordinate system-based approach exemplified already in Subsection 9.2.1 provides a straightforward graphical tool. The *simultaneous display* of different results appears even more frequently in publications. The basis is

usually a two-dimensional ordination upon which the other result is superimposed (as done already in Figures 7.2 and 8.10b). The projection of one OUC over the other eliminates potental disadvantages of either OUC and emphasizes their agreements, as illustrated by the following examples. Graphical evaluation and the superposition of results are warmly recommended in all fields of multivariate data exploration.

First, the 'matrix plot' is used to depict the relationships between a distance matrix and a dendrogram derived from it. Let us choose the dendrogram in Fig. 8.8c, representing the UPGMA clustering of points of Figure 4.3c. The cophenetic correlation (0.662) was already calculated in Subsection 5.5.1. Now, we examine graphically what is behind this correlation. In Fig. 9.19a, the horizontal axis measures the levels in the dendrogram, whereas the distances are measured on the vertical axis. Since the maximum number of different levels in a dendrogram is m–1, the points of the scattergram are arranged in 'columns'. The triangular outline of the point scatter illustrates pretty well that each hierarchial level represents a wide range of original distances, and this range may be especially wide for the last fusions even if the cophenetic correlation is high.

As a confirmation of non-hierachical classifications, the groups may be portrayed by outlines drawn around the member objects in an ordination plane (usually in dimensions 1 and 2). We can do it by eye, but a more elegant solution is to find the minimum *convex hull* (or '*classification polygon'*). This shape includes all points in a group without 'hollows' (i.e., the

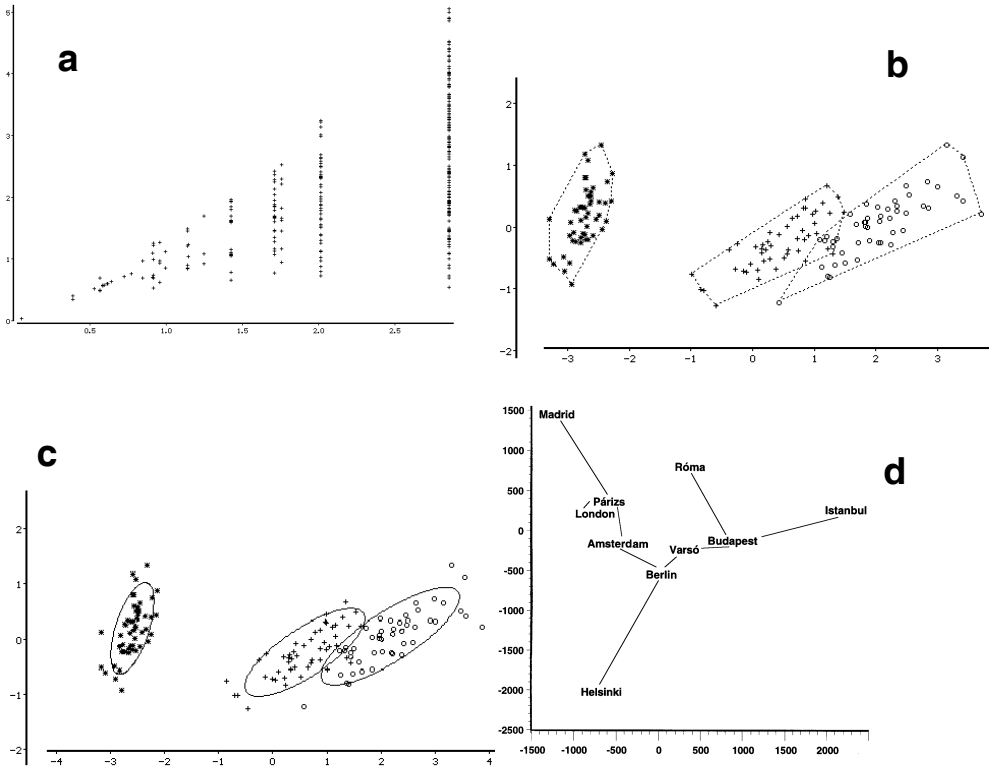

Figure 9.19. Graphical comparison (**a:** dendrogram *versus* distance matrix) and simultaneous illustration (**b:** ordination and a partition with convex hulls, **c:** ordination and a classification with probability ellipses, **d:** ordination and minimum spanning tree) of different types of results

interior angles do not exceed 180^o) such that its area is the minimum. The size of these polygons and the lack of overlaps in between offer a visual basis for evaluating the goodness of a classification. If the overlaps are extensive, the partition is in doubt. As an example, we examine the three-species classification of the *Iris* data drawn on a PCA ordination of the 150 individuals along the first two axes (Figure 9.19b). The diagram is no more than a mere confirmation of the well-established result that one species separates well whereas the other two are not distinct on the basis of the four flower characters included.

Another possibility of contrasting partitions and two-dimensional ordinations is the display of '*ellipses of equal concentration* (Mardia et al. 1979, Lagonegro & Feoli 1985).The ellipse corresponds to an area in the ordination space which contains $100(1-\alpha)$ percent of the members of the given group (α is the probability level chosen). This is only true, however, if the original data follow multivariate normality and sampling is random. These criteria are rarely met in biological investigations. Nevertheless, it is demonstrative to display the probability ellipses of the three *Iris* species at the probability level of 95% (Fig. 9.19c). In this diagram, the relative overlap between ellipses is similar to the classification polygons.

As mentioned already in Subsection 5.4.3, a good visual test of a two-dimensional ordination is offered by *minimum spanning trees*. The closeness of two objects in the ordination plane may be misleading, because the two dimensions portrayed do not represent faithfully enough the interpoint distances. If the edges (links) of the graph cross one another, or there is a long path between two objects that are not far apart in the diagram, then further dimensions should be considered in the ordination display. However, a well-stretched graph without such phenomena is an indicator that the two dimensions are largely sufficient to represent the interpoint distances. This is what we see in the PCoA ordination of European cities (Fig. 7.18): the superimposed minimum spanning tree (Fig. 9.19d) confirms that the 84% share from the total variance is high enough to accept the first two dimensions.

9.6 Literature overview

The literature of the comparison methodology is more extensive than expected and the recent developments in the area make the subject almost impenetrable for an average, yet statistically-minded biologist. This is true of theory only, because the applications to biological problems are unbalanced and often very limited. Data exploration in community ecology, for example, does not exhaust the possibilities in comparison to taxonomy and evolutionary biology. Noted exceptions are the books by Digby & Kempton (1987) and Orlóci (1978). The first book discusses Procrustes methods in detail, whereas the second one devotes much space to the comparison of partitions via information theoretical statistics. The importance of comparisons is clearly recognized by most taxonomic and cladstic monographs, such as Sneath & Sokal (1973). In particular, the review by Rohlf & Sokal (1981a) on the different types and logical pathways of numerical taxonomic comparsions is recommended. By looking at the most recent literature, we find that the search for a consensus cladogram (e.g., Swofford 1991) is the most common preoccupation of evolutionary biologists. We already know why: the number of equally parsimonous optimal cladograms can be exceedingly high for large numbers of taxa.

The need of comparison often arises in a methodological context: of the several alternative methods available one wishes to select the one best satisfying certain predefined basic assumptions. This possibility was not yet mentioned, although there are several examples for this approach, even in ecology (Fasham 1977, Gauch et al. 1977, 1981 etc.). Such studies raise questions like: which ordination procedure is less prone to the arch effect? which method is best suited to recover an assumed (or simulated) background gradient? and so on. Here, the objective is to examine how certain external assumptions are met be the method, rather than to

b

Table 9.2. Comparison of results in different program packages.

	NT-SYS	SYN-TAX	PHYLIP	PAUP
Matrix comparisons	++	++		
Partitions		++		
Dendrograms, trees		++		++
Procrustes methods		++		
Consensus partitions		++		
Consensus trees	++	++	++	++
Significance of comparisons		++		++
Superposition of different reults	++	++		

compare alternative results for the same data. The comparative evaluation of the perfomrance of methods is a different matter.

Although some important sources of informatioin were already cited in the present chapter, it is worth mentioning some journals again whose knowledge is imperative if one wishes to be up to date in a given discipline. Several papers of the *Journal of Classification,* especially the special issue of 3(2) are devoted to the problem of comparing classifications, and understanding of these reviews requires advanced knowledge of discrete mathematics. Perhaps, Rohlf (1974, 1982) and Day (1988) are more suitable as a starting reference. The evolutionary implications of tree comparisons are discussed by Penny et al. (1982, 1991) and, more recently, by Page and Holmes (1998), and almost all issues of *Systematic Biology* offer useful reading for the interested biologist. For ordinations, our choice is more limited.

Regarding the significance of comparisons, the theory and applications of the Mantel test are pioneering and still dominant. Manly (1991) devotes a full chapter to this subject, with ample examples from different biological disciplines. Typical fields of application of the Mantel test are the comparison of phenotypic and genotypic information (Douglas & Endler 1982), comparison of genetic and anthropometric distances (Dietz 1983, whose study relies also on rank correlations), evaluation of point patterns (Harvey et al. 1988) and the elucidation of small scale relationships of species to the environment (Burgman 1987). On the significance of dendrogram and cladogram comparisons, the best reading is Lapointe & Legendre (1990,1991, 1992, but see comments in Podani 2000).

9.6.1 Computer programs

Most commercially available packages quite simply ignore the comparative evaluation of results. There is no excuse for that even though in some cases very special methods are required. Table 9.2 provides a brief list of some software that include routines on pairwise comparisons, consensus, and significance tests.

9.7 Imaginary dialogue

Q: *First of all, let me assure you that – after working through this chapter – most of my scepticism is over. When reading the starting pages I did not understand why is this topic so important for you. Later, at least some of the examples were convincing enough for me to see that*

multivariate data exploration, in most of the cases, does not conclude by generating the results, the OUCs, no matter how attractive they appear for the superficial investigator.

A: Thanks for the recognition! As I mentioned in the literature review, the importance of the topic is largely overlooked in several areas. I tried to resolve this by a review (Podani 1989d) at least in the field of vegetation science, but I have collected no more than 8-10 references since then! Maybe, my efforts will receive more attention later, but it is also possible than scientific fashions will divert people from this topic further apart...

Q: *If I understood well, then the Procrustes method is suitable to ordinations of equal dimensionality only. What shall I do if I am excited to know how an ordination is changed along with increasing dimensionality, i.e., when more and more axes are considered? It is also reasonable to ask how many dimensions of a PCoA ordination fit best to a two-dimensional nonmetric ordination.*

A: Surely, the Procrustes method does not work in these cases. Do not worry, however, because by the good old matrix comparisons you will be able to find the answer. The $m \times m$ distance matrices representing the ordinations may be calculated based on as many dimensions as you wish.

Q: *I would not be surprised to hear about some spatial series analysis associated with these comparisons...*

A: There was already one example, do not you remember? In finding the majority rule consensus, the criterion may be changed from 50 to 100%, thus generating a series of consensus classifications (or other types of results). There is another method that I did not mention yet. Stinebrickner (1984) proposed a family of consensus methods characterized by a modifiable s parameter. If $s = 1$, then we have the strict consensus, while systematic decreases of this value produce more and more clusters in the consensus tree. Of the dendrograms in Fig. 9.16, the Adams tree (g) is identical to the Stinebrickner consensus for $s = 0.5$.

Q: *If there is a proposal to utilize the ultrametic property of the dendrograms in their comparison, then is there any possibility to rely upon the four-point metric in the comparsion of additive trees?*

A: Yes, the '*quartet metric'* (see Steel & Penny 1993, and references therein) implies this for unrooted trees. For each possible quartet of objects (there are " m choose k*"* of them), we examine the two trees being compared. The number of quartets for which the two trees have *different* topology provides a metric distance of trees.

Q: *It seems fairly obvious from what you are saying that consensus trees are of little interest outside cladistics. But why?*

A: Consensus trees are inevitable if the complete hierarchy is of primary concern. Since the evolutionary pathways of a given group of organisms are interesting to the finest detail, cladograms occur most commonly in consensus tree-seeking. In a vegetation study, however, even though full dendrograms are obtained in the first phase of the study, the branching pattern of the tree near the leaves is in fact irrelevant. The partitions that can be derived from the dendrograms at particular high hierarchical levels are more interesting, as illustrated in Figure 9.15. This is so in may other fields of science where the construction of trees is only a first step in a long methodological sequence.

Q: *Is not it potentially dangerous that there may be more consensus methods than the number of trees we are analyzing? May not it be true that the proliferation of methods will overcomplicate our job, leading to an overproduction of results?*

A: This is a proper note, not entirely without irony! Yes, there are many conceivable consensus results, and I can tell you that I did not even mention the majority of consensus methods in this book. There have been long-standing debates over the utility of the consensus approach in biology. I think the problem cannot be circumvented in data exploration, because the number of methods themselves is still steadily increasing. Of course, it is true that a consensus tree could be more appropriate if based on more properties, rather than on a single one (let me remind you of my views on dendrogram comparisons!). Such a 'multivariate consensus' may be of more general validity than the consensus results we already know.

Q: *I am not sure that the consensus should always be searched for in the manner you described. For example, if the alternative trees are based on evaluations of separate subsets of data for the same taxa, then why do not we summarize this information at the level of data, thus saving hours of work with the consensus generator routine?*

A: You should consult some issues of *TREE (Trends in Ecology & Evolution*) in the library nearest to you. There was a debate quite recently in this journal about the possibilities of combining information in phylogenetic reconstruction (vol. 1996, e.g., Ballard 1996). One approach suggests what you have just said: synthesize all possible data first and use this combined set for the generation of the final tree. Proponents of the other approach maintain that it is always interesting to see the alternative evolutionary hypotheses generated by different sets of data, and then to find a compromise among them. Statistical tests may be used to evaluate the null hypothesis that the trees represent the same evolutionary relationships, i.e., their differences are within reasonable limits. If this is true, then *a piori* pooling of data was right. If the null hypothesis is rejected, then the 'many trees' approach is the only one capable of revealing the reasons behind the significant differences among trees.

Q: *If pairwise dissimilarities can be tested for significance, then we should also be able to evaluate a consensus result along similar lines, I think.*

A: A good point again! I can give you an example which answers your question at least partially. Felsenstein (1985) proposed using bootstrap trees in cladistics, each tree being based on a random choice of variables. When a sufficient number of trees are obtained, a majority rule consensus tree is generated. In this tree, each group is examined to see how many percent of the bootstrap trees contained that group. Some clades may have appeared in all bootstrap cladograms, these are the most 'significant' groups in cladistic analysis. Other groups may have lower percentages, but never smaller than 50%. The lower the percentage, the less supported is the given subtree by the data. The percentages are indicated by small numbers on each branch of the consensus tree. Of the many examples for this approach in molecular systematics, I have chosen Krajewski & Dickerman (1990) and Cracraft & Helm-Bychowski (1991) randomly. The recent issues of the journal, *Molecular Phylogenetics and Evolution* provide a large number of case studies in which bootstrapping plays a central role. However, there is no unanimous enthusiasm about bootstrapping and subsequent consensus generation

among phylogeneticists. A common criticism is that estimates regarding the significance of clades are too conservative (Hillis & Bull 1993).