8

Matrix rearrangement

(How to do without mathematical constructs?)

The methods discussed thus far convert biological information into mathematical objects, such as dendrograms, cladograms, ternary plots, unrooted trees, components, factors, and so on, used in order to express the essence of the data in comprehensive form. Admittedly, the interpretation and understanding of results conveyed by these 'artificial' constructs require experience and some basic knowledge of theory. Our life would be easier if the final results were a direct and more obvious illustration of data structure for everyone. To achieve this goal, let us forget about these mathematical objects by forcing the end result to be of the same type as the starting data! For such an approach, the data matrices themselves should deserve the most attention. Their rows and columns can be rearranged in such a way that mere inspection of the 'new' matrix reveals the hidden structure in the data. The rearrangement of distance or similarity matrices serves similar purposes. This chapter will treat procedures that are suitable to such intuitively meaningful reordering, with or without doing preliminary analyses by other methods known from the previous chapters. Of course, the description of the algorithmic details requires some mathematics even though the final results speak for themselves. Logically, this chapter could have been presented much earlier in this book, but I think that elementary knowledge of classification and ordination theory is very helpful at several points. First, reordering of variables in a data matrix will be discussed and then follow procedures which rearrange both the columns and the rows, implicitly achieving ordination or classification objectives.

8.1 The unequal importance of variables: character ranking

The sequence of variables in data matrices prepared by hand is usually accidental or it follows some practical rule, such as the alphabetic order of names. This is not critical at all for multivariate analysis, because the results should always be independent of the input order of data (if this requirement is not met, then there are serious problems with the method itself or with the software, see Podani 1997c). Nevertheless, one may want to obtain an objective order

of variables so as to reflect their importance in determining the structure of data. Such a list should start with variables that dominate the data structure, followed by the less important ones and the list should be concluded by variables whose removal from the data is almost unnoticeable. The keystone in this approach is in fact the definition of *importance*. As we shall see below, there are several, more or less equally meaningful definitions. Fortunately enough, importance can be quantified in objective manner based on a wide selection of formulae. Furthermore, ranking variables also depends on whether it is performed before more sophisticated analyses (*a priori* ranking) or its purpose is simply to evaluate the importance of characters in contributing to the outcome of a particular classification or ordination (*a posteriori* ranking). The latter possibility is closely related to the topic of evaluating final results, touched upon already in Subsection 5.5.3 and to be discussed in more detail in Chapter 9.

In addition to the rearrangement of data, there are further objectives of character ranking. For example, Dale et al. (1986) mention the following:

- The selection of most important variables to be retained in subsequent analyses because the program cannot handle all variables simultaneously. This problem is no longer acute if we consider the rapidly increasing performance of microcomputers and computer program packages.
- Simplification of complex, multivariate situations into the univariate case (e.g., the discriminant functions allow a multivariate separation of groups, whereas the dichotomous identification keys use one criterion variable in each step).
- The identification of irrelevant variables that do not reflect biological pattern in any meaningful way. These variables usually produce only background noise, so that their omission from the data may improve the final results of data analysis.

8.1.1 Ranking variables a priori

The objective of *a priori* ranking is to order the variables according to their individual contributions to the entire data structure. The measurement of this contribution depends primarily on the scale on which the original data are expressed. For interval-scale and ratio-scale variables, the covariance-, correlation- or sometimes the cross-products matrices (Equations 3.68-70) serve as a basis for ranking. For binary (presence/absence) data, there are additional possibilities, such as the information theory measures and the χ^2-statistic. These latter two provide a solution for the nominal data type as well. There is another choice that the user must make in advance: this is between the two fundamental strategies of ranking: a) analysis with calculating and removing the residuals in each step, and b) simple ranking without considering residuals.

Ranking by residuals. This method requires several iteration steps. First, the most important variable is chosen and then its effect is removed from the data (Orlóci 1973, 1978). As a result, the residual variation in the remaining matrix is linearly independent from the firstly selected variable. After eliminating the effect of the first variable, the second most important variable is chosen, its effect removed, and so on. The iterations stop when the residual variation becomes zero. This is achieved inevitably for the penultimate variable, but ranking may happen

to be complete earlier when 100% of the variation in the data structure is explained by fewer than n-2 variables. In such cases, the remaining variables cannot be ranked.

This technique is shown first on the example of interval-scale variables, using the $\mathbf{S}_{nxn} = \{s_{jk}\}$ matrix of their cross products, covariances or correlations. As we shall see, the basis of ranking is the contribution to the total sum of squares in the raw, centered or standardized data, respectively. The main steps are as follows:

1. The first rank is $r = 1$. Calculate the quantity, $\mathrm{tr}\{\mathbf{S}\}$, which is in fact the total sum of squares (for cross products) or the total variance (for covariance or correlation) in the data.

2. Determine for each column j of **S** the sum of squared elements and divide it by s_{jj}. The first rank is given to the variable with the highest score. Formally, we maximize the quantity given by

$$g_j = \sum_{k=1}^{n} s_{jk}^2 \,/\, s_{jj} \,. \tag{8.1}$$

Let this variable be denoted by h in the forthcoming steps. Its percentage contribution to the total variation is obtained as $100 \times g_h / \mathrm{tr}\{\mathbf{S}\}$.

3. The effect of variable h is now removed from **S**. Any element of the matrix, including the diagonal values, is recalculated according to

$$s_{jk} = s_{jk} - \frac{s_{jh}\, s_{kh}}{s_{hh}} \,. \tag{8.2}$$

As a result of these operations, all values will be zero in row and column h of **S.** The decrease of all other elements will be proportional to the correlation (covariance, cross-product) of the given variable with variable h.

4. Set $r = r$+1. If there are positive values in **S**, the analysis goes back to step 2. Otherwise, the total sum of squares (variance) is exhausted by the variables already ranked and the analysis terminates.

The above ranking procedure determines the minimum number of *original* variables necessary to explain the total variation in the data. A geometric illustration may facilitate a deeper understanding of the method. The variables are to be conceived as points in an m-dimensional space. Now, s_{jj} is the squared length of the vector directed to point j (scalar product of the vector, see Appendix C), and $\mathrm{tr}\{\mathbf{S}\}$ is the sum of all squared vectors (sum of squares). Each variable is now viewed as if the vector pertaining to it were an axis. There exists a hyperplane perpendicular to each axis, to which the vectors of all variables can be projected. The differences between the original and projected vector lengths appear in the numerator of Formula 8.1. The most important variable is thus the one providing the maximum decrease of sum of squares achieved by this projection, including the variable's own projection to zero length. This is shown in Figure 8.1a for the case of $m = n = 2$. Figure 8.1b demonstrates that one of two fully (actually negatively) correlated variables is 'superfluous' (i.e., the corre-

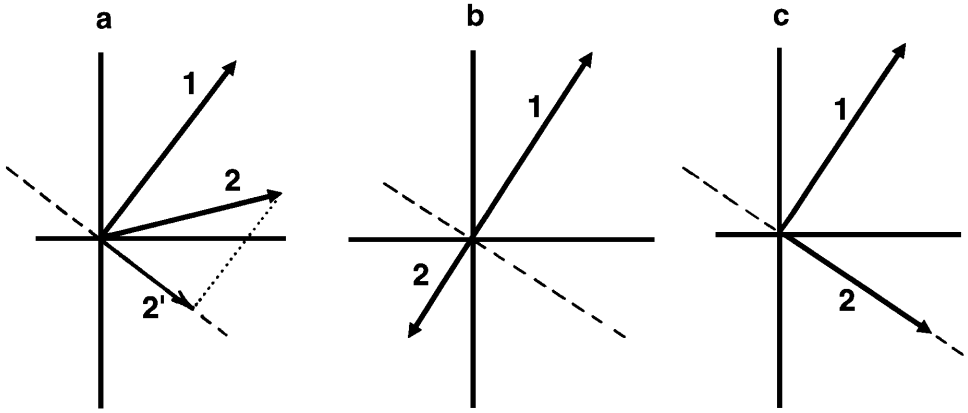


Figure 8.1. Ranking by residuals: a geometric illustration of a single analytical step. In all cases, variable 1 is selected. Then, **a**: the share of variable 2 linearly independent of the first variable is proportional to the length of the vector as projected to the dashed line (2'), **b:** variable 2 has no independent contribution to the sum of squares, and **c:** the two vectors are orthogonal to each other.

sponding vectors are collinear). Whenever the variables are orthogonal to one another – as opposed to the previous situation – none of them explains any variance pertaining to the other (Fig. 8.1c). After identifying the most important variable, the dimensionality of the system decreases by one, and a new variable is searched for in the new subspace.

To sum up: the essence of the method is the decomposition of the total sum of squares into orthogonal constituents. However, this operation does not introduce new artifical constructs (such as components or factors) so that the cumulative percentage accounted for by the first p ordered variables is usually much lower than the cumulative percentage of the first p principal components extracted from the same data. These percentages would be equal only if the original variables were coincident with the components, a situation almost impossible in practice. The advantage of using the original variables over components is that they are directly interpretable for the biologist.

As an example, let us examine the variables of Table A1 based on the three criteria introduced above, The results are summarized in Table 8.1. For cross products, eight variables were necessary to reach 100%, for the other two criteria seven were sufficient because the variables were centered. The rank of the starting matrix (Appendix C) strongly influences the number of variables that can be ranked[1]. The result is a bit surprising for the cross products matrix (Table 8.1A) because the firstly selected variable has a relatively small sum of squares (18.0), whereas others have much higher (e.g., BRO ERE has 3020.0 and SES SAD has 4916.0). This demonstrates elegantly that it is the directionality represented by the species vector in the multidimensional space that matters, rather than the absolute variation. The axis coincident with CAR HUM accounts for 41.9% of the total sum of squares; an amount higher than for any other species. Of course, for the covariance criterion the rank order is greatly different (Table 8.1B), showing the effect of data centring. Here, species with high variance dominate, whereas species with low variance, such as CAR HUM do not even appear in the rank order. As a result of standardization (Table 8.1C) – as expected – the rank order is

1 The two meanings of the term 'rank' should not be confused

Table 8.1. Ranking the species of Table A1 using the strategy of eliminating residuals. The species remaining after 100% of the variance was reached are omitted. Small inconsistencies are due to rounding errors.

	Rank	Variable	Specific share	Relative importance	Cumulative %
A	1	CAR HUM	5297.278	41.935	41.935
Cross products	2	SES LEU	3629.493	28.733	70.668
	3	BRO ERE	2656.635	21.031	91.699
	4	CHR GRY	549.148	4.347	96.046
	5	FUM PRO	284.417	2.252	98.298
	6	SCA CAN	123.509	0.978	99.275
	7	CAM SIB	50.065	0.396	99.672
	8	SES SAD	41.487	0.328	100.000
		Total:	12632.000	100.000	
B	1	SES SAD	651.905	53.642	53.642
Covariance	2	BRO ERE	318.132	26.178	79.820
	3	SES LEU	161.852	13.318	93.138
	4	CHR GRY	59.445	4.891	98.029
	5	FES PAL	18.822	1.549	99.578
	6	SCA CAN	4.483	0.369	99.947
	7	KOE CRI	0.647	0.053	100.000
		Total:	1215.286	100.000	
C	1	CAR LIP	4.061	33.840	33.840
Correlation	2	FUM PRO	2.372	19.763	53.603
	3	CHR GRY	1.961	16.345	69.949
	4	SES SAD	1.576	13.131	83.080
	5	SES LEU	0.951	7.925	91.004
	6	BRO ERE	0.882	7.346	98.350
	7	FES PAL	0.198	1.650	100.000
		Total:	12.000	100.000	

changed again. A typical feature of correlation-based rank orders is that the cumulative percentages (last column) increase much more slowly than in the previous two orderings.

One may ask the question: under what circumstances is this particular ranking strategy preferable? The answer is simple: in any case when the multivariate method to be applied in the same survey reduces dimensionality of the data on the same theoretical grounds as used in ranking. In other words, all the methods should be logically compatible; a principle further clarified below with examples. Otherwise, the rank order and the results cannot be contrasted.

After ranking, the size of the data matrix can be reduced considerably without strongly modifying the results of compatible multivariate analyses. For example, a centred PCA restricted to the first three species of the rank order (93%, Table 8.1B) gives practically the same result for the first two components as the entire set of species (the reader may verify this agreement easily). Also, a standardized PCA does not change much if the least important species in the correlation-based ordering (Table 8.1C) are omitted. However, there is no point in using the rank order where the strategy is logically incompatible with the subsequnt analytical method, as is the case with removing residuals before classifications. The logic of clustering

requires that *simple* ranking (see below) be used for selecting the best subset of variables. In fact, ranking by residuals cannot serve as a basis for tabular rearrangement, because the omission of some variables, no matter how negligible they are, appears to be some sort of information loss for the researcher! This method is the only one not recommended for rearranging the rows of data tables.

> For presence/absence data, Orlóci (1976a) suggested measuring the contribution of variables to their mutual information (Formula 3.115). The rank order is established by finding in each step the variable whose deletion causes the maximum decrease of mutual information. The logic behind this choice is that this variable is the most informative on the other m-1 variables in the data matrix. After its removal, the second most important variable is found, and so on. Inevitably, there is a tie in the last two positions of the order, but the mutual information may fall to zero much earlier. A disadvantage of the method is its high computing demand. Expansion of the formula facilitates the analysis of multistate nominal variables. In addition to information theory measures, some derivatives of the χ^2 statistic may also be applied to 2^m contingency tables (Fienberg 1970).

Simple ranking. In numerical classification, either hierarchical or non-hierarchical, the highly correlated variables have a 'synergistic' relationship, one corroborates the effect of the other. It is generally acknowledged that the more variables support a classification, the more general is its validity. Therefore, it would be unwise to eliminate any variable that correlates with another variable already selected in a previous step of the ranking procedure. A different ranking principle should be adapted in order to show the absolute contribution of the variables to some overall measure. In this case, the decomposition is not orthogonal, all the variables may be ranked and the the entire data table may be rearranged. When defining a ranking criterion, one may think of the variance of variables, assuming that variables with low variance are likely to be much less influential to the separation of classes than those with high variance. (It is a different matter that any variable may prove to be useful *a posteriori*, but recall subsection 5.5.3.) Such a ranking procedure has been used implicitly when rare species are deleted from extremely large ecological data tables. The cross products, covariances and correlations among variables may also be used without considering the residuals. This means that the values obtained in step 2 of the algorithm of p. 287 are used as the basis of ranking and the analysis stops here (Podani 1994). If we recall the interpretation of Figure 8.1, then it becomes clear that in this way the importance of each variable is a measure of how that variable represents the others in the m-dimensional space. Therefore, variables with odd behaviour or causing mere stochastic noise in the data get to the end of the rank order. The importance values of variables may be added, and thus the percentage contribution of a group of variables to the total may also be calculated. This is useful to evaluate the relative importance of variable groups in the data matrix.

> Let us examine the species of Table A1 using the procedure of simple ranking. The first two columns of Table 8.2 reflect proportions in the total variance, but these figures should not be taken very seriously. This variance, as shown by the example of CAR HUM, is uninformative as to the relationships among variables. A species with relatively low performance values may turn out to be very important in simple ranking. The rankings based on cross products and covariances are more similar to each other than in Table 8.1. Such a ranking is recommended in any case when the absolute quantities are considered important (i.e., when a subsequent classification uses Euclidean distances or sum of squares). The rank order derived from correlations is preferred whenever the data are standardized. It is useful to prepare a rear-

Table 8.2. Simple ranking of the species of Table A1 based on four criteria.

	Variance			Cross products			Covariance			Correlation		
	Species		%	Species		%	Species		%	Species		%
1	SES SAD	604.50	49.74	CAR HUM	5297.2	12.1	SES SAD	651.9	16.9	CAR LIP	4.0	10.8
2	BRO ERE	280.28	23.06	SES SAD	5212.1	11.9	CAR HUM	594.9	15.4	SES LEU	3.8	10.3
3	SES LEU	119.69	9.84	FES PAL	3840.8	8.8	FES PAL	374.2	9.7	FES PAL	3.8	10.2
4	CHR GRY	104.12	8.56	SES LEU	3723.0	8.5	SES LEU	363.7	9.4	CAM SIB	3.3	8.8
5	FES PAL	69.14	5.68	KOE CRI	3673.3	8.4	BRO ERE	322.8	8.3	CAR HUM	3.2	8.7
6	SCA CAN	16.26	1.33	CAM SIB	3673.1	8.4	CAR LIP	297.5	7.7	SES SAD	3.2	8.7
7	FUM PRO	13.92	1.14	BRO ERE	3626.1	8.3	CAM SIB	289.2	7.5	CHR GRY	3.1	8.5
8	KOE CRI	2.26	0.18	CEN SAD	3407.1	7.8	CHR GRY	227.6	5.9	KOE CRI	2.6	7.2
9	CAR HUM	1.92	0.15	SCA CAN	3024.2	6.9	FUM PRO	207.7	5.4	BRO ERE	2.5	6.8
10	CEN SAD	1.42	0.11	CAR LIP	2964.5	6.8	SCA CAN	200.6	5.2	CEN SAD	2.5	6.8
11	CAR LIP	1.14	0.09	CHR GRY	2796.7	6.4	CEN SAD	167.7	4.3	FUM PRO	2.4	6.4
12	CAM SIB	0.57	0.04	FUM PRO	2311.3	5.3	KOE CRI	148.3	3.8	SCA CAN	2.3	6.2

ranged table in which the variables are listed in their order of importance. For the covariances in our small example, this rearrangement will be given by

SES SAD	0	0	0	0	0	0	4	70
CAR HUM	1	0	0	0	0	0	1	4
FES PAL	20	11	5	15	25	4	6	2
SES LEU	25	15	0	8	25	1	1	0
BRO ERE	5	7	18	0	1	0	50	11
CAR LIP	2	0	1	1	3	1	0	0
CAM SIB	0	1	0	0	0	0	2	1
CHR GRY	30	8	5	0	4	0	0	0
FUM PRO	3	11	7	5	7	12	3	2
SCA CAN	1	10	0	0	0	0	2	8
CEN SAD	1	1	1	4	1	2	3	3
KOE CRI	5	1	2	1	1	0	2	1

In the first rows of this new table, we find the species most 'responsible' for the data structure. One is warned, however, not to consider the remaining variables as being absolutely useless in clustering, because low ranked variables may be decisive in affecting minor details of classifications.

There have been many alternatives to simple ranking. The multiple correlation of each variable (a special case of canonical correlation with one variable in the first group and n-1 variables in the other, see Subsection 7.2) is a reasonable criterion which is worth mentioning. Rohlf (1977) and Orlóci (1978) discuss this procedure in detail and point to the disadvantage that the computational demand of this method is high.

Dale & Williams (1978) consider the entire data matrix as a contingency table (as in COA), and suggest to calculate the expected value of each cell based on the row and column totals, as done in calculating the χ^2 statistic (denominator of Formula 3.36). The sum of the absolute values of these differences for each variable ("eident value") is the basis of ranking. A version of this strategy anologous to ranking with residuals involves determining the most important variable and recalculating the contingency table in each step of the analysis.

In the pioneering age of numercial classification, the ranking of binary (presence/absence) variables (species) utilized the χ^2 statistic as part of monothetic divisive clustering (Subsec-

tion 5.3.2). The interspecific association matrix was computed in each clustering step, and its column totals were used in ranking (Formula 5.7). The variable with the maximum column total was considered to be the most informative with respect to the others. Formula 5.8 appears more suitable to this *a priori* ranking because it is insensitive to small cell frequencies.

8.1.2 A posteriori ranking

Determining the importance of variables in the results, potentially followed by the rearrangement of raw data, should be the integral part of almost all methods of multivariate analysis. This point is not new in this book; I already mentioned possibilities for evaluating the results of hierarchical classification (Subsection 5.5.3). There are many more opportunities for *a posteriori* ranking, of course, to be discussed below in a framework that follows the main categorization of multivariate methods. Again, one has to bear in mind that the ranking criterion should be compatible logically with the distances or other resemblance measures and data transformation methods used previously to derive the particular result being evaluated.

Importance of variables in partitions. In the optimality criterion of the k-means procedure (J, Formula 4.1), the influence of variables is additive (summation according to i). Therefore, partitioning of J into components attributable to individual variables and subsequent ranking of these contributions pose no problems for us. A variable that explains a given classification perfectly has a 0 contribution to J; this happens if the variable has a constant value within each class. On the other hand, the variables not supporting or even conflicting with the actual partition will account for most of the variation expressed by J. In the global optimization strategy, the role of variables is much less explicit. First, dissimilarities are calculated and then their averages are computed during the classification process, therefore it is more difficult to trace the effect of individual variables. The general procedure introduced at the end of Subsection 5.5.3 was developed for this complicated situation. The Ψ_{ik} measure reflects the extent to which variable i contributes to the within-cluster distances (or dissimilarities) relative to its contribution to between-cluster distances (dissimilarities) for k classes. (For the calculation of these contributions, see Podani 1997b). Using the Ψ measure, the variables can be ranked such that the ordering is compatible with the resemblance function used in the classification. In fuzzy clustering, the contribution of variables to the fuzzy sum of squares may be easily obtained using Formulae 4.6 and 4.7. Then, arranging these contributions in ascending order will provide the ranking of variables.

In the fuzzy classification of the three *Iris* species with the coefficient of fuzziness f = 1.25 (Fig. 4.9), the variable contributions are as follows: PW 10.7%, SW 19.7 %, SL 33.2 % and PL 36.4%. This is not surprising because raw data were used in the calculations and the mean values of variables increase in that order. For f = 2.5, the situation does not change much, although the two length characters are interchanged so that the length of sepals will be the most contradictory with the classification.

There are some further possibilities for *a posteriori* ranking of variables. The ratio of the between-group variance to the within-group variance (if >0) of a variable, the F-statistic, was proposed by Jancey (1979). The ratio of between-group variance to the total variance was used by Lance & Williams (1977). Those authors used a contingency table for each binary or nominal variable (rows are classes and columns the character states in this table) and then expressed the discriminative power of the variable by the Cramér index (Formula 3.37).

Hierarchical classifications. A hierarchical classification can be conceived as a series of partitions so that the role of variables may be evaluated for each level separately using any method described above (a typical example is the method proposed by Lance & Williams, 1977). A variable that was found to be extremely important in the two-cluster classification of objects may be in conflict with the classification into three or more classes, while other variables may exhibit the opposite behavior. For this reason, there is no point is seeking the *global* importance of variables in hierarchies.

Cladograms. The importance of characters in a phylogenetic hypothesis is readily evaluated using the consistency index (6.9) or the retention index (6.11). Characters entirely supporting a given cladogram take the value of 1 and, in a rearranged data matrix, they should appear in the first rows, followed by characters with decreasing values of these indices. It is likely that the ordering is only partial, due to tied values.

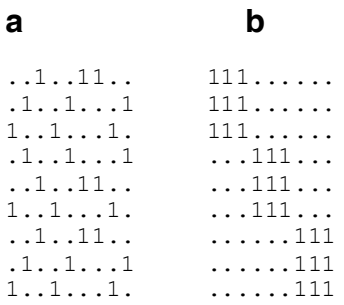
Importance of variables in ordinations. The ranking criterion may be selected in many different ways, depending primarily on the ordination method itself. Since ordinations are most commonly portrayed in two dimensions, the most interesting question to ask is to what extent a given variable corresponds to the arrangement of points along axes 1 and 2. In principal components analysis, one possibility of *a posteriori* ranking is to measure how much percent of the variance of the character is explained by the two components, as calculated by Formula 7.12.

> Summation of the percentages in the lower part of the first two columns of Table 7.1 reveals that the ordination of Fig. 7.2 best reflects the following species: SES SAD (99%), BRO ERE (87.7%), CAR HUM (86.8%), SES LEU (63.8%), FES PAL (62.5%), CAM SIB (61.6%) and CAR LIP (54.5%) whereas the most conflicting species is KOE CRI (6.2%). This ordering corresponds quite closely with the simple *a priori* ranking determined using covariances (Table 8.2). Omission of KOE CRI from the data therefore would have a negligible effect upon the ordination result.

For standardized PCA, ranking may follow the same strategy as above. The sum of squared correlations between a variable and two selected components provides a measure how strongly the two components explain that variable. Recall that a variable gives a unit sum of squared correlations with *all* the components! In *canonical correlation analysis*, we can use Functions 7.26-27 for ordering the variables; separately for the the two domains, of course. In *correspondence-analysis,* the distance from the position of a variable to the origin is a measure of importance. The larger this distance, the more influential is the variable on the arrangement of objects. As in standardized PCA, the unimportant variables are positioned near the origin. In *multidimensional scaling*, variables cannot be ranked because the original data are not used. If they are nevertheless available, then their correlations with the axes may be considered for ranking. After *canonical variates analysis*, the communality of variables (Formula 7.79) serves as a basis for ranking, as demonstrated already in Table 7.2.

Rearranged tables. Although matrix rearrangement as a specific multivariate procedure of its own was not discussed yet, the importance of variables in rearranged matrices must be mentioned for completeness. The relative contribution of variables to rearranged data matrices

Figure 8.2. A completely disordered matrix (**a**) may obscure the strong interaction between rows and columns, visible only after block-clustering (**b**). For clarity, 0-s are replaced by dots.



produced by block clustering (8.2.3) or block seriation (8.2.4) may be used for ranking the variables, as well as the objects[3]. For block clusters, the procedure follows the logic of jackknifing: the criterion of block-sharpness is determined with and without the given variable, and the difference between the two values provides the desired measure. For the χ^2 criterion, this difference may be both positive and negative: the latter result (decrease of χ^2 when the variable is removed) implies that the presence of the variable intensifies block structure. Positive difference indicates that the variable interferes with the data blocks and its deletion would lead to a more sharply structured matrix. Thus, the rank order begins with the variable with the highest negative value. When block sharpness is measured by entropy or sum of squares, the removal of a variable cannot cause negative changes. In this case, the most important variables are those with the smallest contribution to the criterion used. When the diagonal structure is optimized in the data matrix (Section 8.3), the variable contributions are additive and may be calculated easily by Formula 8.10. The greater the contribution, the more uncertain is the position of the variable in the rearranged matrix.

8.2 Block clustering

In the previous subsection, I raised the possibility of rearranging data matrices based on the variables, but this operation would be incomplete without considering the objects as well. If the variables and the objects alike are classifiable into meaningful clusters, then the rearranged matrix should reflect these groupings. Such a rearrangement has a great interpretive value: groups of variables explain the partition of objects and vice versa. The classification of rows and columns divides the data matrix into small submatrices or blocks, each reflecting the relationship between certain groups of variables and objects. For presence/absence data, such a relationship is the most clear-cut if certain blocks contain only 1-s while others only 0-s. The inherent block structure of the matrix is not seen in a data table with rows and columns written in an arbitrary order: the exploration of this hidden information is the objective of block clustering algorithms. The problem is illustrated by the deliberately simple matrix of Figure 8.2.

The exploration of hidden block structure of data matrices may be needed in several fields of science. In biology, for example, the rearrangement of very large ecological tables has been

3 In this case, the objective is not the preparation of a new rearranged table, as previously, because rearrangement is already given. Ranking variables , however, is still useful here for a better interpretation of data blocks.

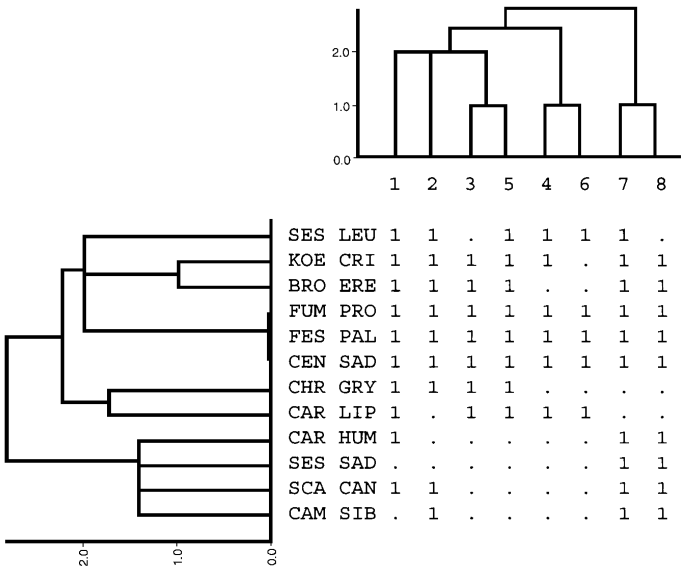


Figure 8.3. Block-clustering of the presence/absence version of Table A1, with groups obtained by complete linkage clustering from Euclidean distances. The figure illustrates the presence of ties, often observed for small binary data matrices (the ties were resolved by single linkage resolution, cf. Section 5.2). The delineation of blocks and the order of objects and variables are arbitrary to some extent.

the main objective of the Zürich-Montpellier phytosociological school since its beginnings (cf. Braun-Blanquet 1965, Mueller-Dombois & Ellenberg 1974). Manipulation by hand, however, is a cumbersome task with questionable results. The development of computers has changed this situation completely. As the most straightforward solution, one may classify the variables as well as the objects using the same clustering method, and the groups resulting from the two separate analyses may be used to reorganize the table. The first such study is apparently due to Williams & Lambert (1961a,b). In agreement with the attribute duality principle, the strategy of association analysis (Subsection 5.3.2) was applied to plant ecological quadrats ("normal association analysis") by summing up χ^2 values for the species, and then to species ("inverse analysis") by considering χ^2 sums for the quadrats. The data matrix was rearranged according to the groups obtained by 'cutting' the dendrograms at arbitrarily selected levels. The method has been known as "*nodal analysis*" (Greig-Smith, 1983, reviews similar, computer-oriented procedures developed later in numerical plant ecology). The divisive method can be replaced by other hierarchical or non-hierarchical methods that are applicable to both variables and objects. The "projection" of the two classifications onto each other will provide the required rearrangement. It is hoped that the interplay between classes of variables and objects is best revealed this way (Figure 8.3).

This is not always the case, however. The two classifications, even though derived from the same data matrix, are in some sense "independent" from each other. When groups of columns are formed, the program ignores the classifiability of the columns and vice versa: the classification of columns does not assume any grouping of the rows. The *interaction* among object and variable groups is best revealed when the blocks directly arise from a heuristic search or an optimization algorithm (Gordon 1981). Separate row and column cluster analyses cannot achieve this goal, so that methods not yet introduced in this book are needed. Thus, this subsection can be conceived as a late continuation of the chapters on classification.

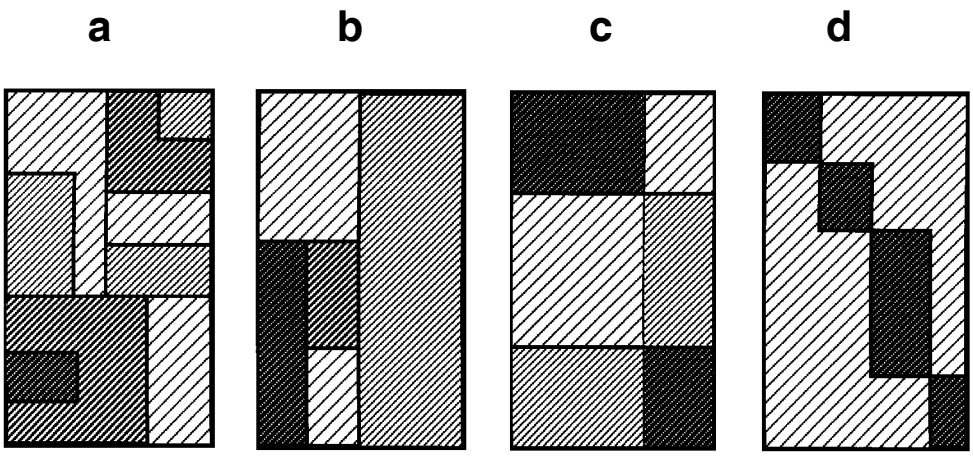


Figure 8.4. Main objectives of block clustering. **a:** Unconstrained blocks, **b**: partial block clustering. **c**: cross-partition, general case ($p \surd q$), **d**: block-seration ($p = q$). Shading reflects within-block homogeneity.

Methods of block-clustering are categorized into four main groups according to the constraints applied during the classification of rows and columns:

- In the simplest case (*unconstrained block clustering*), there is no grouping at all for rows or columns; the objective of rearrangement is to disclose maximally homogeneous blocks or clusters of values within the matrix. The data blocks may take irregular shapes (Figure 8.4a).
- In *partial block-clustering*, the rows are classified into p, the rows into q classes, but any row-wise class may characterize two or more groups of columns, and vice versa. The data blocks are rectangles (Figure 8.4b).
- In *cross-partitions* or fully blocked data matrices, any data value may belong to only one row group and one column group (Figure 8.4c). It is allowed, although not required that $p \neq q$.
- In the special case of $p = q$, one may impose a one-to-one correspondence between groups of rows and groups of columns in the matrix (Figure 8.4d). This is called *block seriation*, a procedure transitional towards the methods to be discussed in Section 8.3. In this case, emphasis is placed upon the diagonal blocks, usually entirely disregarding the arrangement of values falling outside.

8.2.1 Unconstrained block clustering

Hartigan (1975) lists several algorithms developed for unconstrained block arrangement. One such technique, the "*two-way joining*" strategy applies to binary data. The complement of the simple matching coefficient (3.6) is used as a dissimilarity coefficient to compare rows by rows and columns by columns. In each step of the analysis, the mutually nearest two rows or two columns are moved right besides each other in the matrix. The number of maximally homogeneous groups is determined automatically during the analysis. By adapting a threshold

	1	7	8	2	3	5	4	6
BRO ERE	1	1	1	1	1	1	.	.
CEN SAD	1	1	1	1	1	1	1	1
FES PAL	1	1	1	1	1	1	1	1
FUM PRO	1	1	1	1	1	1	1	1
KOE CRI	1	1	1	1	1	1	1	.
SES LEU	1	1	.	1	.	1	1	1
CHR GRY	1	.	.	1	1	1	.	.
SCA CAN	1	1	1	1	.	.	.	.
CAM SIB	.	1	1	1	.	.	.	.
CAR HUM	1	1	1	.	.	.	.	.
SES SAD	.	1	1	.	.	.	.	.
CAR LIP	1	.	.	.	1	1	1	1

Figure 8.5. Two-way joining for the binarized form of Table A1. Zeros are replaced by dots for clarity.

value for within-block homogeneity, the method can be modified to conform with interval scale variables.

The result of two-way joining of the presence/absence version of Table A1 is shown in Figure 8.5. The algorithm forms as many groups as minimally required to include only zeros or 1-s. Therefore, quite many blocks occur in the final result, and their delineation is not always equivocal. The difference from Figure 8.3 is substantial. It is easy to see from this example that for large data matrices the result can be confusing and difficult to interpret.

Hartigan's (1981) another procedure is suitable for evaluating categorical data. The initial elements of blocks are selected by the *leader* algorithm (cf. Subsection 4.1.4) based on the criterion that their distances from the other elements exceed a prespecified threshold. The very first element is the upper left value (x_{11}) of the matrix. The algorithmic steps alternate between rows and columns. If too many groups arise, each with a single value only, the threshold is too low and the analysis should be repeated with a higher distance value.

For the presence/absence case, Bruelheide & Flintrop (1994) suggest using a threshold as well: each block is formed by variables that appear in at least ε percent of the objects of the same block, and vice versa. The algorithm forms the blocks with stepwise elimination of rows and columns. The resulting matrix often has no clear-cut block structure, because the elements of a given block may be isolated by some other blocks (see their Table 8). Eckes (1995) attempts to minimize the increase of sum of squares through the agglomerative "*centroid effect method*". The relatively complicated algorithm is in fact an adaptation of the hierarchical clustering method utilizing criterion 5.5. According to the author, the final blocks are obtained by stopping the fusions when the increase of sum of squares is "excessive". That is, the stopping rule is fairly arbitrary and a more objective criterion is sought.

8.2.2 Partial block-clustering of data matrices

Gordon (1981) mentions several procedures designed for this purpose, with emphasis on Hartigan's (1972) divisive procedure. This accepts data measured on the interval or the ratio scale, because the within-block sum of squares are minimized when concentrating the most similar values in the same block. The algorithm operates as follows. At the outset, there is no

specific requirement as to the number of blocks. Let $\bar{z}_{ij}$ denote the mean value of the block to which data value x_{ij} belongs, and then the goal is to minimize the quantity:

$$J = \sum_{i=1}^{n} \sum_{j=1}^{m} (x_{ij} - \bar{z}_{ij})^2. \tag{8.3}$$

Hartigan (1972) proposed searching the minimum by an hierarchical strategy. The data matrix, and in the latter steps the blocks are successively divided into two parts either by rows or by columns, depending on whether the row-wise or column-wise division provides the maximum decrase of the J measure. However, the result depends greatly on the initial ordering of the rows and the columns, so that the method should be modified to generate permutations as well to allow other starting situations. Without this, the divisive method can only be used if the order of rows and columns is unequivocal, that is, defined by some meaningful external criterion or a ranking procedure. Dale & Anderson (1973) proposed to divide the data matrix in a similar way, using a monothetic criterion.

8.2.3 Cross-partitions

The objective is to arrange the variables into p, and the objects into q classes such that the resulting cross-partition, that is, the $p \times q$ rectangular blocks of the matrix satisfy some optimality criterion. Podani & Feoli (1991) selected three such criteria from the many possibilities for measuring *block sharpness*:

- sum of squares within blocks for interval and ratio scale variables (Formula 8.3, denoted this time by $J_{(p,q)}$);
- weighted within-block entropy for nominal characters:

$$H_{(p,q)} = \sum_{i=1}^{p} \sum_{j=1}^{q} \left(k_i k_j \log k_i k_j - \sum_{h=1}^{s} f_{hij} \log f_{hij} \right) \tag{8.4}$$

  where k_i is the number of elements in variable group i, k_j is the number of elements in object group j, s is the number of states of the nominal characters ($s \geq 2$) and f_{hij} is the frequency of character state h in block ij;
- the blocks are considered as the cells of a $p \times q$ contingency table, and the sum of values within each block ij is taken as the cell frequency (f_{ij}). Then, Formula 3.36, denoted here by $\chi^2_{(p,q)}$ applies. This formula is suited to presence/absence data, although its formal application to data comprising frequencies (such as counts) is also conceivable.

The strategy is to minimize the first two criteria or maximize the third in order to obtain maximally homogeneous blocks. The experienced reader may immediately see that this is a hard problem, because for fixed values of n, m, p and q the possible number of rearrangements, as computed by the Stirling formula (4.17) is $S_{(n,p)}S_{(m,q)}$, which is an astronomical number for most practical problem sizes (within-block ordering is immaterial, of course). In lieu of algorithms that lead to the optimum in reasonable time, we restore to heuristic searching strategies. The method proposed by Podani & Feoli (1991) is an iterative algorithm which relocates a row or a column of the matrix such that the criterion is maximally improved in each step. The

a

	1	2	7	8	3	4	5	6
CAM SIB	.	1	1	1	.	.	.	.
CAR HUM	1	.	1	1	.	.	.	.
SCA CAN	1	1	1	1	.	.	.	.
SES SAD	.	.	1	1	.	.	.	.
BRO ERE	1	1	1	1	1	.	1	.
CAR LIP	1	.	.	.	1	1	1	1
CEN SAD	1	1	1	1	1	1	1	1
CHR GRY	1	1	.	.	1	.	1	.
FES PAL	1	1	1	1	1	1	1	1
FUM PRO	1	1	1	1	1	1	1	1
KOE CRI	1	1	1	1	1	1	1	.
SES LEU	1	1	1	.	1	1	1	.

b

	1	2	7	8	3	4	5	6
CEN SAD	1	1	1	1	1	1	1	1
FES PAL	1	1	1	1	1	1	1	1
FUM PRO	1	1	1	1	1	1	1	1
KOE CRI	1	1	1	1	1	1	1	.
BRO ERE	1	1	1	1	1	.	1	.
CAM SIB	.	1	1	1	.	.	.	.
CAR HUM	1	.	1	1	.	.	.	.
CAR LIP	1	.	.	.	1	1	1	1
CHR GRY	1	1	.	.	1	.	1	.
SCA CAN	1	1	1	1	.	.	.	.
SES LEU	1	1	1	.	.	1	1	1
SES SAD	.	.	1	1	.	.	.	.

Figure 8.6. Most optimal block clustering of the presence/absence version of Table A1 based on the χ^2 and the J statistics (**a**) and entropy, H (**b**) after 100 iterations such that $p = q = 2$.

iterations stop if there are no rows or columns the relocation of which would cause further improvement. For the $J_{(p,q)}$ criterion, this method is the two-way generalization of k-means clustering. Being an iterative method, the final result is greatly influenced by the starting configuration (order of rows and columns) and – depending on data structure – the iterations may easily converge into suboptimal solutions. There is no guarantee that even from hundreds of random initial arrangements the analysis will converge into the (absolute) optimum, although the best of the results may be quite close to it. The relatively large computing demand of the procedure is not a serious problem nowadays, even if the number of rows and columns exceeds 100.

Binary data have the obvious advantage that all the three criteria discussed above apply to them. With parameters $p = q = 2$, 100 runs were performed on the presence/absence version of Table A1 for each criterion. After selecting the best one from each series, the rearranged matrices of Figure 8.6 were obtained. One interpretation of the results is as follows.

Using the χ^2 statistic, the iterations produced one extremely bad result ($\chi^2 = 1.75$), the maximum value of χ^2 (10.1) resulted 42 times (Figure 8.6a), while there were many outcomes with intermediate success. For the J measure, the same value ($J_{max} = 12.75$) and the same rearrangement were obtained in all the 100 analyses! In case of entropy, the partition of sites was the same in all results, whereas the best species classification varied (Figure 8.6b, $H_{max} = 223.57$, occurring 62 times). It is to be pointed out that suboptimal values (224.76 and 225.54) were also obtained in a large number of iterations with the only differences between the classification of columns. These examples demonstrate sufficiently that the three criteria do not necessarily lead to the same final configuration and if so, their efficiency in finding the optimum is considerably different. The block containing only 0-s in Fig. 8.6a, or other two blocks containing almost only 1-s in Fig. 8.6b cannot appear simultaneously in the result, but our combined approach detected all of them.

It may have a great interpretive value to identify variables that explain the rearrangement most perfectly or, on the other hand, to find those that strongly conflict with the result. As mentioned already in Subsection 8.1.2, the variable contributions are readily expressed by the percentage change after the removal of the given variable from the data. The order below lists the variables from the most explanatory to the most conflicting:

	Δchi²%		ΔJ%		ΔH%
.1	SCA CAN -27,08	1	FUM PRO -2,52	1	SES SAD -5,83
2	CAM SIB -20,31	2	FES PAL -2,52	2	CAR HUM -7,24
3	CAR HUM -20,31	3	CEN SAD -2,52	3	CAM SIB -7,24
4	SES SAD -13,54	4	SCA CAN -2,61	4	SCA CAN -8,69
5	CAR LIP -11,38	5	CAR HUM -5,88	5	CHR GRY -9,31
6	CEN SAD -2,88	6	CAM SIB -5,88	6	CAR LIP -11,54
7	FES PAL -2,88	7	KOE CRI -7,28	7	BRO ERE -12,17
8	FUM PRO -2,88	8	SES SAD -10,46	8	SES LEU -12,49
9	SES LEU -2,16	9	SES LEU -12,04	9	KOE CRI -13,01
10	CHR GRY -1,44	10	BRO ERE -12,61	10	FUM PRO -13,01
11	KOE CRI 1,55	11	CAR LIP -18,49	11	FES PAL -13,01
12	BRO ERE 6,42	12	CHR GRY -22,69	12	CEN SAD -13,01

Although the rearranged matrix itself is identical for the first two criteria, there are differences in the rank order of variables. As expected, the variable effects differ with the measure of block sharpness. For example, CAR LIP is strongly discriminative between the two groups so that is removal would result in a significant decrease of χ^2; it contributes much to the sum of squares of the lower left block of the table. It is left to the reader to inspect further details of the results. The new order of objects is also worth examining:

	Δchi²%		ΔJ%		ΔH%
.1	8 -25,92	1	5 -2,94	1	4 -11,57
2	7 -23,08	2	3 -7,19	2	6 -11,57
3	5 -18,27	3	4 -12,09	3	3 -13,24
4	3 -15,90	4	1 -13,40	4	5 -14,94
5	4 -13,70	5	7 -14,71	5	2 -17,79
6	6 -11,42	6	2 -17,65	6	8 -17,79
7	2 6,00	7	6 -17,65	7	1 -19,14
8	1 11,72	8	8 -20,26	8	7 -19,14

For reasons detailed above, the sequence of objects is not identical in the first two cases, notwithstanding that the optimal block structure is the same.

Constrained block clustering. In the chapters on clustering and ordination, we saw already some methods which force the analysis to run within certain limits. Block clustering may also be constrained, for example, by keeping the partition of either the rows or the columns fixed. This may be necessary in ecology, when the classification of sample sites is not allowed to change because it is already a result of a consensus analysis of many alternative partitions (see Section 9.4) and we need the best explanatory clusters of species. In this case, only the species are relocated from one cluster to the other during the iterations. The reverse procedure is also possible; the partition of variables is fixed in order to find the best groups of objects to optimize, say, an identification key.

Concentration analysis. Having finished the block clustering of presence/absence data, the mutual interpretability of row and column classes may be enhanced by ordination ("*analysis of concentration*", Feoli & Orlóci 1979). This is in fact a symmetrically weighted correspondence analysis of groups (Section 7.3) after the adjustment of within-block sums, f_{ij}, based on the formula:

$$F_{ij} = \frac{f_{..} f_{ij}}{n_{ij}} \Big/ \sum_{g=1}^{p} \sum_{h=1}^{q} \frac{f_{gh}}{n_{gh}} . \qquad (8.5)$$

In this, F_{ij} is the new value and n_{ij} is the size of block ij. Adjustment is necessary to eliminate excessive differences in block sizes, so that all of them will attain equal importance (Orlóci & Kenkel 1985). The number of possible ordination axes is $t = \min\{p-1, q-1\}$. The χ^2 -statistic calculated from adjusted blocks is different from the optimized value in the iterations and is useful to measure the goodness of rearrangement *("relative divergence"):*

$$RD = \frac{\chi^2}{t\, f_{..}}. \tag{8.6}$$

RD ranges from 0 to 1, indicating the sharpness of block structure relative to the possible minimum and maximum. The *RD* score may be used to find the best block structure in the most general case where the values of p and q are allowed to change.

8.2.4 Block-seriation

The methods introduced in the previous subsection consider only the internal homogeneity of blocks; the ordering of row and column groups is free. The most highly specialized methods of block clustering go a little further by attempting to maximize the contrast between blocks in the diagonal position and the others (Figures 8.2 and 8.4d), so that ordering becomes crucial. Such approaches require that $p = q$. Whereas in cross-partitions all blocks are equal in importance, block seriation[4] (Marcotorchino 1991) emphasizes diagonal structure and treats the non-diagonal blocks as a single unit, irrespective of the internal structure. Block seriation is most commonly applied to presence/absence data. The p-block seriation of binary matrix **X**, such that a row group is denoted by A_k, a column group by B_k, involves maximizing the Garcia - Proth (1985) criterion given by the fomula:

$$GP_p = \sum_{k=1}^{p} \sum_{i \to A_k,\, j \to B_k} x_{ij} \;+\; \sum_{k=1}^{p} \sum_{i \not\to A_k,\, j \not\to B_k} (1 - x_{ij}). \tag{8.7}$$

Its meaning is perhaps simpler in words than in mathematical formalism: the goal is to concentrate as many 1-s into the diagonal blocks, leaving as few as possible in the off-diagonal blocks. In a perfect situation, the diagonal blocks contain only 1-s, the others 0-s, causing GP_p to reach its maximum value, nm. Consequently, GP_p/nm may be used as a relative measure of diagonal block sharpness, with values falling within the interval [0,1]. When optimizing this index, we encounter the same problems as in many occasions before: exact algorithm is known only for relatively small n, m (<30) (Marcotorchino 1991), but optimizing *GP* for real problem sizes is impossible within reasonable time. It may happen that the heuristic methods of the previous sections give a good approximation to the optimum value of *GP*, or even hit it, but there is no guarantee for its perfect success.

> In plant ecology, block seriation is an attractive procedure where the communities change along a single, background gradient and yet, there are well-distinguishable vegetation types or noda. The starting data are not always of the presence/absence type; but counts or cover percentages cannot be used for optimizing the Garcia-Proth criterion[5]. Two-way indicator spe-

4 The meaning of the word "seriation" will be clarified in the next section.

5 Nevertheless, coefficient (8.7) can be modified to comply with any non-ordinal data types to measure the discrepancy between diagonal and off-diagonal blocks.

cies analysis (Hill 1979a), as incorporated into the **TWINSPAN** program (Subsection 5.3.1) may be a good solution of this data-type problem, although it does not have any explicit optimality criterion. The analysis is based on the simultaneous ordination of species and sites, followed by their two-cluster partitioning according to the ordination scores on each axis. The end-result of this combined ordination-classification approach is a diagonally structured matrix in which, if there is a close agreement between species and site groups, the blocks are easily recognized (there is no p value to be specified in advance). If there is no inherent group/block structure in the data, then the problem reduces to simple seriation by reciprocal averaging, with potentially meaningless divisions (but see the next subsection). Wildi (1989) suggested a more complex interaction of classification and ordination algorithms with the aim to eliminate noisy elements (either species or sites) from the data and to maximize recognition of diagonal blocks in this manner.

Block seriation of distance matrices. The discussion of block-seriation may be expanded readily to distance or similarity matrices. Now, the objective is to concentrate small distances (large similarities) into the diagonal blocks to maximize contrast with the other, off-diagonal blocks. The rearranged matrix will reflect the classification of objects; the more clear-cut the groups, the sharper is the block structure. Since row i is the same as column i, the strategy will be simpler than for data matrices: the relocation of a row will automatically involve relocation of the corresponding column. This problem rarely appears alone; matrix rearrangement is usually based on previous classifications and serves as an *a posteriori* illustrative vehicle to clarify an existing classification.

The block seriation of distance matrices is closely related to the classical problem of matrix shading: the cells of the matrix are colored such that their darkness is proportional to the distance values [6]. The simplest, and most commonly used strategy is to generate first a hierarchical classification of objects and to reorder the objects in the distance matrix according to the dendrogram (Figure 8.7). However, a dendrogram can be drawn in 2^{m-1} ways without changing the classification topology itself (recall Subsection 5.1, Figure 5.2), so that the definition of blocks is inevitably arbitrary to some extent. Gale et al. (1984) have shown that finding the best dendrogram shape is associated with the conventional (non-block) seriation of the distance matrix, so that it is time to turn our attention to the next section.

8.3 Seriation

Seriation involves finding a simultaneous ordering (permutation) of the rows and the columns of data matrices with the objective of revealing background one-dimensional gradients. The basic idea is that large scores should be concentrated along the diagonal, while the low values should fall as far from it as possible. Seriation was first used for the indirect dating of geological strata based on archaeological and paleontological findings. In other words, seriation facilitates reconstruction of a temporal sequence of objects based on the continuous change of their characteristics (Kendall 1970, 1971, Goldmann 1971). In biology, time is just one factor to consider in explaining a sequence; an ecological gradient may also be responsible for the simultaneous ordering of objects and variables. It is now clear that seriation is a special

6 The shaded blocks in Figure 8.4 rely upon our intuitive feelings that in fact any block clustering result can be illustrated by matrix shading. This opportunity did not raise in case of Figs 8.5-6 because shading is not needed for presence/absence data.

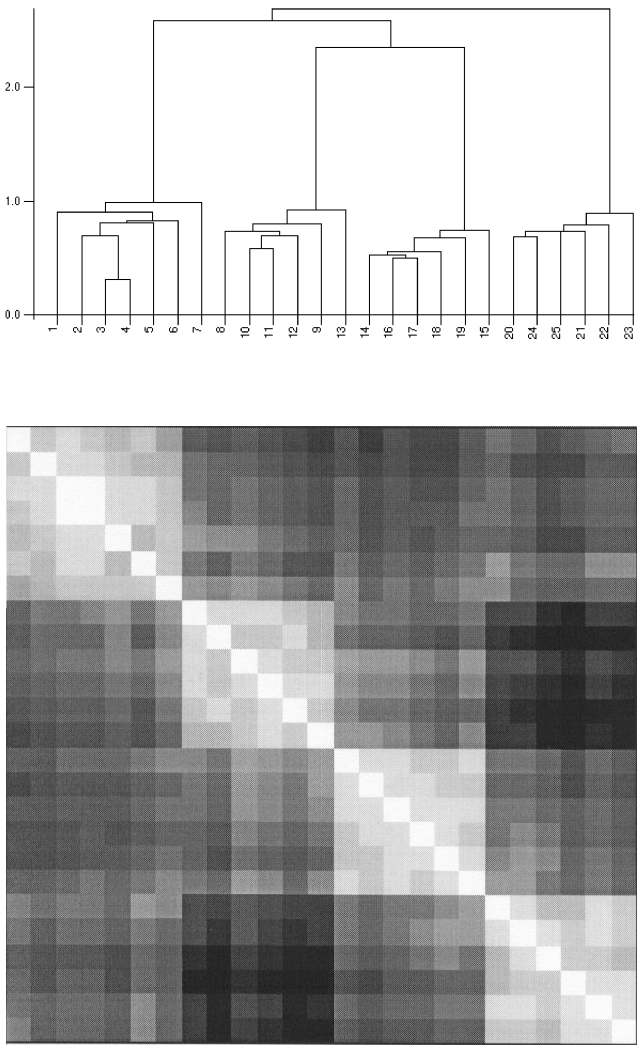


Figure 8.7. Rearrangement and shading of the distance matrix for the points of Figure 4.3 after the single linkage clustering of objects (Fig. 5.6b). Usually, the congruence between a clustering and the matrix is less clear than in this example. The diagram was pre-pared using the **SYN-TAX** Mac pro-gram; white corresponds to the minimum distance (0), black to the maximum (7.87) with a continuous transtition between the extremes.

one-dimensional ordination applied directly to the raw data, so that this section could be considered as a continuation of Chapter 7. Distance matrices may also be seriated, of course. In this case, a permutation of objects is sought so that small distances are concentrated near the diagonal (for similarities, the opposite is the goal). This problem is simpler than data seriation, because only one permutation is optimized. Using the terminology of matrix shading, the objective is to find a dark diagonal band in a data or similarity matrix, which is continuously lightened towards the lower left and upper right corners (or the other way, for distances).

8.3.1 Seriation of data matrices

The rows and the columns of a data matrix may be ordered numerically in many ways, for example, by ordination. Since reciprocal averaging (correspondence analysis, Subsections 7.3.1-2) is essentially a simultaneous ordination of objects and variables, it appears to be best suited to this problem if the data comprise frequencies. Reciprocal averaging is the basis of the classificatory program **TWINSPAN** as well. Principal components analysis may also be considered for this purpose, because the objects and variable scores on the first axis may provide a meaningful ordering for matrix rearrangement. An ordination-based reordering is most efficient if a strong one-dimensional background gradient dominates in the data; in other words, the first eigenvalue is relatively large.

> The largest eigenvalue from the correspondence analysis of dune vegetation data (Table A4) is 0.53 (25%), indicating a 'moderately strong' background gradient. The reordering of rows and columns based on the coordinates on the first axis is shown in Figure 8.8a. The non-zero values (in gray) are arranged in a distinct, wide band along the diagonal. The species present in most sites occur in the middle of the table. The identification of the background gradient requires further analyses; the CCOA using the same data and external information on environmental variables (Fig. 7.17) is the most logical choice for this purpose.

Ordination-based seriations rely upon one (usually the first) axis, and reflect sequential information only for the particular dimension selected. Clearly, the other, linearly independent components of the variance are disregarded in this rearrangement. Consequently, if the first

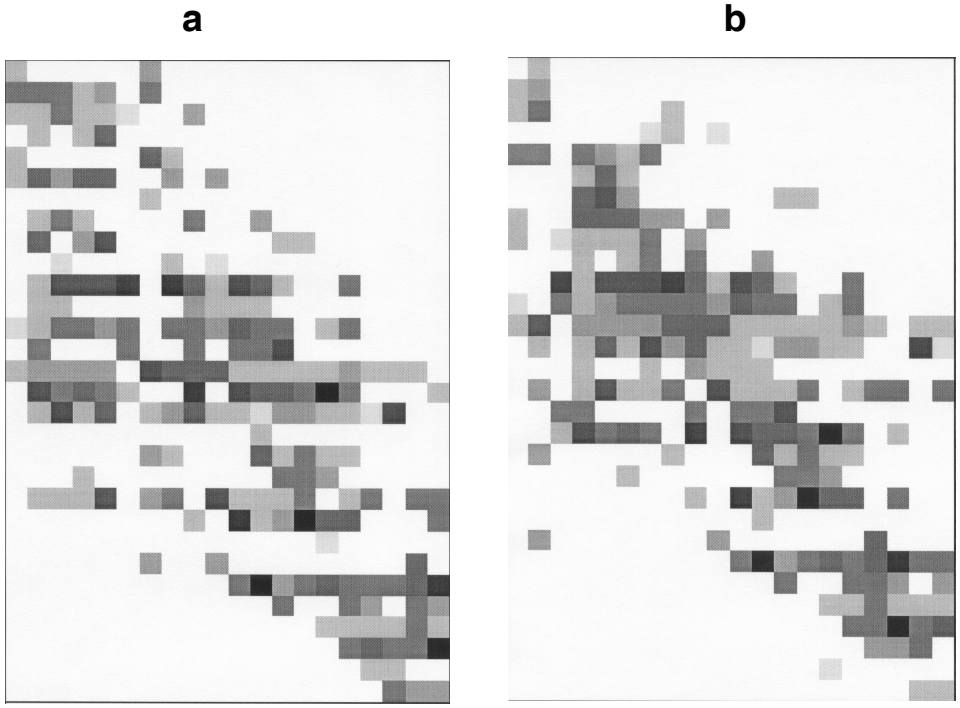


Figure 8.8. Seriation of the dune vegetation data (Table A4) according to the first correspondence analysis axis (**a**) and by maximizing criterion 8.10 (**b**).

eigenvalue is relatively small, the rearranged matrix will not be very informative. In fact, ordinations do not optimize any function measuring the goodness of rearrangement directly, and we cannot expect that they will provide results comparable to direct seriation methods. The direct procedures use a criterion variable referring to the diagonal arrangement of values, explicitly considering only a one-dimensional sequence for both the rows and the columns. The task is enormous, because the number of possible sequences is exactly $n!m!/4$ (the number of row permutations multiplied by the number of column permutations, then divided by 2×2, because the direction of orderings is immaterial for us). For large values of n and m, there is no possibility to try all the permutations, and no algorithms are available yet for finding the optimum in reasonable time.

McCormick et al. (1972) proposed maximizing the *neighborhood criterion* for the permutations of n rows and m columns:

$$MC_{P(n),P(m)} = \sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij} \, (x_{ij+1} + x_{ij-1} + x_{i+1j} + x_{i-1j}) . \qquad (8.8)$$

In this, $x_{0j} = x_{n+1j} = x_{i0} = x_{im+1} = 0$ by definition. Formula 8.8 concentrates upon the local conditions in the data matrix, and cannot be expected to provide acceptable results under all circumstances. In order to find a more general solution to this problem, we should see that Formula 8.8 can be decomposed into the sum of two terms:

$$MC_{P(n),P(m)} = \sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij} \, (x_{ij+1} + x_{ij-1}) + \sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij} \, (x_{i+1j} + x_{i-1j}) = MC_{P(m)} + MC_{P(n)} , \qquad (8.9)$$

that is, into components by columns and by rows. It has the consequence that the rows and the columns can be treated separately and that the matrix can reordered using the two optimal permutations. The authors propose a heuristic solution of the famous 'traveling salesman' problem to find the optimum of (8.9).

> It is sufficient to examine how the method works for rows because the same algorithm applies to the columns. The basic idea is to take row 1 as the pivot element, and then all other rows are examined whether they should be inserted *before* or *after* the pivot row in order to maximize the increase of $MC_{P(n)}$. After relocating the row that provides this maximum, we examine the remaining n-2 rows each of which can be placed into 3 positions. Then, the optimal position of the remaining n–3 rows is selected from the 4 possibilities, and the iterations continue in a similar way until all rows take their new position. Since the result depends on the choice of the pivot element, the whole procedure is repeated n-1 times to allow all rows to become the first element. Then, the best of the n results is accepted as the final permutation, even though we cannot be sure that this is the absolute (global) optimum for row rearrangement. The search is repeated for the columns, and then the matrix is reordered according to the best two permutations.

Other functions consider more than the immediate neighborhood of diagonal values. The most attractive feature of rearranged matrices is the so-called Robinson property (Robinson 1951). A matrix is said to possess this property if its values decrease monotonically in the rows as well as in the columns when moving away from the diagonal in both directions. For data matrices, the Robinson property is implicitly considered by the following criterion (Podani 1994):

$$\Psi_{(n,m)} = \sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij} \left[\left| \frac{m \square i}{n} - j \right| + \left| \frac{n \square j}{m} - i \right| \right]. \qquad (8.10)$$

In this formulation, each x_{ij} value is weighted by the sum of its positional differences from the diagonal. These are the number of rows (within column j) and the number of columns (within row i) that value x_{ij} should be moved to get right into the diagonal. Usually, this is not an integer. Thus, if a large value falls far from the diagonal, then its contribution to measure 8.10 will also be large. The objective is therefore to minimize the amount of the Ψ measure in some way. The minimum implies that the matrix is in the closest state to meet the Robinson condition. Since the sum of 8.10 cannot be decomposed into row and column contributions – contrary to Formula 8.8 – separate optimizations of row and column permutations cannot help. Similarly to the algorithm of block-clustering (8.2.3), a random initial configuration may be modified iteratively by relocating the row or the column that provides the best improvement in each step. This is time-consuming for large matrices. Also, there is a high chance that the iterations converge into local optima so that many parallel runs are necessary, with no guarantee that the best result will ever be found. The very poor local optima often arising from random configurations can be avoided if the starting permutations rely upon a previous ordination (e.g., RA), as illustrated below.

For the sake of comparison with RA reordering, the iterative procedure is applied to the vegetation data of Table A4. From 50 random initial permutations, the best result was $\Psi = 5078$. The reordered and colored matrix is shown in Figure 8.8b. According to criterion 8.10, this is superior to the RA-based rearrangement (Fig. 8.8a), for which $\Psi = 5698$. However, the RA-result was considerably improved by iterations and the goodness of rearrangement ($\Psi = 5093$) strongly approximated the best randomization-based result. (To be true, most of the 50 randomization-based analyses produced a better outcome than RA.) A visual comparison of the two colored matrices of Figure 8.8 shows unequivocally that the use of the Ψ measure produces a *higher concentration* of large values along the diagonal than RA. As a 'compromise', some small values get much further away from the diagonal than in the RA rearrangement. The degree of difference between the two results is always case dependent, so that it is worth trying both procedures.

8.3.2 Seriation of distance and similarity matrices

This is an easier job than optimizing data matrices, because we have to worry about only one permutation. The rearrangement may be based on an ordination, as before. Measure 8.10 also applies here, so much the more because its formula and its optimization simplify greatly for symmetric matrices:

$$\Psi_{(m)} = \sum_{i=1}^{m-1} \sum_{j=1}^{m} s_{ij} \, | i - j |. \qquad (8.11)$$

Similarity matrices are optimized by the algorithm applied to raw data using criterion 8.10, whereas for distance matrices the value of Ψ is minimized.

One hundred random arrangements of the distance matrix of European cities (Table A4) were analyzed by the iterative relocation algorithm of seriation. Since matrix size is relatively small, the best result ($\Psi = 354400$) was reached in as many as 64 runs (Fig. 8.9a). The colored matrix demonstrates the frequent case where we cannot remove all the large distances from near the diagonal. Helsinki, Madrid and Istanbul are at the tips of a triangle and their distance relationships cannot be portrayed in one dimension sufficiently. The appearance of large val-

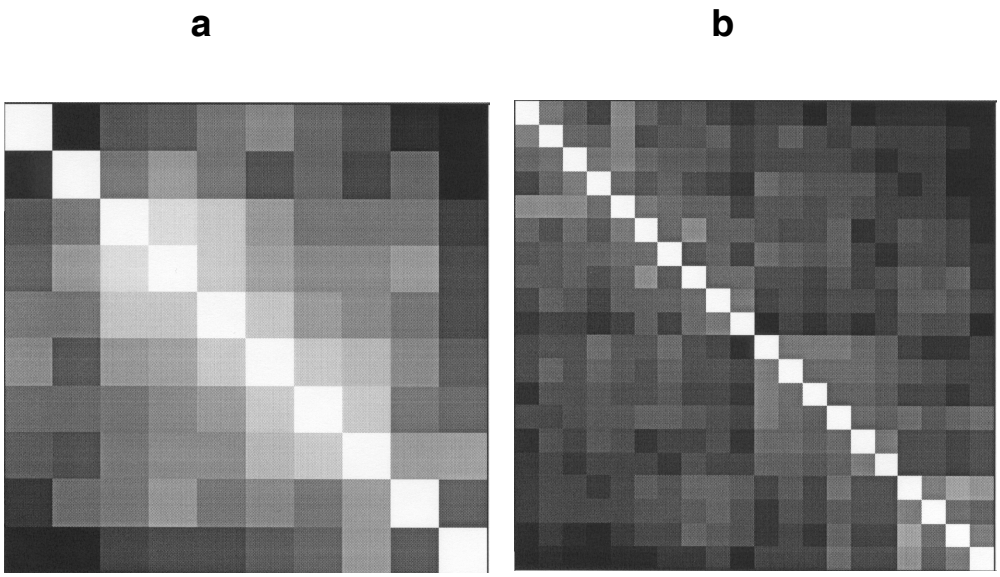


Figure 8.9. Seriation of the Euclidean distance matrix of European cities (Table A7, **a**) and of the columns (sample sites) of Table A4 (**b**) by maximizing the quantity 8.11.

ues in the corner of the matrix indicates a situation analogous to the horseshoe effect often observed in ordinations. Seriation was somewhat more 'successful' from the Euclidean distances for the objects (columns) of Table A4 (Ψ = 18410; Figure 8.9b). This optimum occurred 14 times out of the 50 runs of the algorithm. The order of objects disagrees with the sequence seen in Figure 8.8 and two relatively distant objects (3 and 19) are too close to each other in the middle of the diagram, showing further similarity to the horseshoe effect. The arrangement along the diagonal can be used for a visual detection of blocks, suggesting a two-cluster grouping of objects with a potential subdivision in the left group. Recall the proposition put forward by Gale et al. (1984): seriation may be used indirectly to recognize group structure in the data.

The Robinson condition applies directly to resemblance matrices, as shown by Hubert et al. (1982). They suggested to create a standard Robinson matrix in which one element $x_{ij} = m-|i-j|$. Thus, all values in the diagonal are equal to m, and the values decrease regularly and monotonously off the diagonal. The most optimal seriation of a similarity matrix is therefore the one showing maximum positive correlation with the Robinson matrix. For distances, the negative correlation is to be maximized.

8.4 Literature overview

The literature of character ranking and matrix rearragement methods is much poorer than usual in other fields of multivariate analysis. I have made references to most of the relevant papers and books already. For character ranking, the best source of information is Orlóci (1978). Most of the algorithms developed for finding internal blocks in data matrices are introduced by

Table 8.3. Character ranking and matrix rearrangement in some program packages.

	Statistica	BMDP	SYN-TAX
Ranking by elimination			++
Simple ranking			++
Two-way joining	++		
Block seeking ("leader algorithm")		++	
Block clustering by iterative relocations			++
Concentration analysis			++
Iterative seriation from distance and data matrices			+

Hartigan (1975). In the biological sciences, the need of tabular rearrangement appears most commonly in plant ecology and phytosociology, therefore most textbooks devoted to this subject have several references to the manual and automated procedures. Marcotorchino (1991) and the references therein are a good starting point for those interested in the mathematical background of seriation algorithms. An early and still useful summary of the topic is the book edited by Hodson et al. (1971). In addition to the key-paper by Kendall, this volume comprises four other contributions to the theory and application of archaeological seriation. Matrix shading is treated in detail by McIntosh (1978), mainly from an ecological and phytosociological viewpoint.

8.4.1 Computer programs

There are relatively few program packages that offer solutions to ranking and reordering problems. Table 8.3 provides a summary of relevant features available in packages already mentioned in this book. The routines in **SYN-TAX** (Podani 1989c) save the rearranged matrix according to the new ordering of variables. Shaded matrices are also part of the output, for both block-clustering and seriation (the figures in this chapter were also prepared by this program). **Statistica** uses several colors applied to categories of data values to an attractive display of the results of two-way joining.

Programs written in the BASIC language are listed for character ranking and concentration analysis in Orlóci (1978) and Orlóci & Kenkel (1985). The best-known program for the phytosociogy-oriented audience is undoubtedly **TWINSPAN**, whose code (Hill 1979a) appears to be 'transmitted' to several other program packages (e.g., **PC-ORD**, McCune 1986). Program **TABORD** (Maarel et al. 1978) also deserves mention, although I am not aware of any upgrades of it to current systems.

8.5 Imaginary dialogue

Q: *You admitted that some ranking methods cannot be used for rearranging data matrices, then why do you discuss them in this chapter?*

A: As I mentioned, if the ranking procedure cannot provide a full ordering of variables, then the remaining variables have to be provided in an arbitrary sequence at the end of the new data matrix. Nevertheless, I maintain that all ranking procedures should be treated in the same place, and this chapter is the most suitable to this purpose. I am convinced that in published data matrices the variables should be presented in order of their importance, unless there is a

strong reason to do it otherwise (e.g., in alphabetic order, or species grouped according to life forms).

Q: *The rigorous division of ranking procedures into a priori and a posteriori methods excludes the possibility of intermediate strategies...*

A: I see what you mean: you imagine an iterative algorithm in which each step evaluates the importance of characters and gives higher weight to the important ones before refining the ordering in the subsequent computations. The iterations could stop when we have a sufficiently stable result. Actually, Jancey & Wells (1987) already proposed a realization of this idea in a classificatory context. In each step of the divisive algorithm, there is a re-ranking of variables to ensure that each variable is emphasized at that hierarchical level where it is most important, whereas at other levels it is disregarded as being a source of noise. Fowlkes et al. (1988) provide an overview of iterative refinement procedures, although they propose selecting a subset of most relevant characters, rather than ordering. This built-in variable screening is called *forward selection* in the literature.

Q: *I like your proposal that the variables most responsible for, or most conflicting with a given block-structure are selected by removals, one at a time. Could you apply the very same strategy to evaluate the importance of variables in ordinations as well?*

A: This looks a reasonable suggestion. An ordering of variables based on their influence upon a particular result could be achieved, for example, by comparing a reference ordination (the one based on all the variables) to the others, each of which obtained by deleting a single variable. If the similarity to the reference ordination is high, then the variable is less important, because after its omission from the data almost the same result is reproduced. One the other hand, if the removal of a variable decreases the similarity to the reference ordination considerably, then it was essential in affecting the configuration. The crucial step in this procedure is the way similarity of ordinations is assessed and I recommend to wait with this until the relevant methods are treated in the next chapter.

Q: *If it is true, then you contradict yourself! On the analogue of the above proposition, the global influence of variables upon dendrograms is also a measurable quantity, contrary your statement in Section 1 where you insist that the effect of variables should be evaluated level by level in a hierarchy.*

A: This is a very good point I did not think of earlier! I agree that the reference dendrogram could be compared with n other dendrograms, each generated by the removal of a given variable. Then, the rank order of variables is obtained according to these n pairwise similarities. And there are plenty of methods for the comparison of dendrograms, as you will see in the next chapter. Of course, this procedure is time-consuming if n is high and this explains that, as far as I know, nobody has ever attempted to do such evaluations. Good idea for a master's thesis, for example!

Q: *Why do you reject concentration analysis if the data are not of the binary type?*

A: In principle, the analysis has the same assumptions about the starting data as correspondence analysis or, in general, the χ^2-statistic. Namely, the data must be presences and absences, or frequencies. For other kinds of data, this function is simply meaningless. The block-clustering program will run with any data type and provide meaningful result, but sub-

sequent correspondence analysis of blocks, again, is not recommended at all if the above mentioned conditions are not satisfied.

Q: *I suspect that the traveling salesman algorithm you mentioned is not much faster than complete enumeration (exhaustive search of all possibilities).*

A: If you wish, we can check... After selecting a given row, the other n-1 rows may be assigned to two positions, and then there are three places for n-2 rows, and so on until only one row remains for which we have exactly n possibilities. These terms sum up, and this sum must be multiplied by n because all the rows may be pivot elements. After doing the same for the columns, you can add the two sums to get the total number of steps necessary to complete the analysis:

$$n \sum_{i=2}^{n} i\,(n-i+1) + m \sum_{j=2}^{m} j\,(m-j+1). \qquad (8.12)$$

This sum is certainly less than the number of possible orderings. For $n = 8$ and $m = 7$, we have that $n!m!/4 = 50\,803\,200$, whereas quantity 8.12 equals only 1275. This sufficiently illustrates that the traveling salesman algorithm examines a small fraction of the possible arrangements and therefore it has a high chance to miss the global optimum. For real problems, such as $n = 100$ and $m = 80$, the difference between complete enumeration and heuristic search is even more drastically different.

Q: *How about the order of importance of objects in seriation? You did not mention that such a ranking is possible to assess how the individual objects support a given rearrangement.*

A: Yes, I simply forgot it... However, by looking at Formulae 8.10 and 8.11 one sees immediately that the contribution of any row or column to Ψ is readily obtained and then ranking is a matter of arranging these values into descending order.

Q: *Coloring and shading distance and data matrices are artistic activities in my opinion, even though the computer does these for us. I think, Piet Mondrian would accept some diagrams as representatives of his school. I heard of colored graphs as well and I would like to know if they are also interesting for the biologist.*

A: Colored graphs provide further means of disclosing hidden information in the data. Similarly to tabular rearrangement, their interpretation does not require much mathematical background. In colored graphs, termed *plexus graphs* in ecology, the vertices represent the study objects or variables (species). The edges or lines of the graph are colored or are drawn with different thickness depending on the distance, similarity or correlation between the objects they represent. By convention, increasing similarities are visualized by thicker or darker edges. A plexus graph is therefore an alternative to shaded $n \times n$ or $m \times m$ matrices and it is not surprising that McIntosh (1978) discussed both groups of methods in the same review. The plexus diagram is easy to draw for a few points (Fig. 8.10a), but as the number of vertices increases, the problem of their efficient arrangement on the plane becomes more complicated. Two-dimensional ordinations do help in this regard (e.g., Matthews 1978, Matus & Tóthmérész 1990). Figure 8.10b is the plexus graph for the European cities drawn using their ordination shown in Fig. 7.18. As you see, many edges are missing from the graph: in fact, large distances are visualized by deliberately deleting the edges pertaining to them. The plexus method is a useful auxiliary tool for evaluating ordination diagrams, as demonstrated

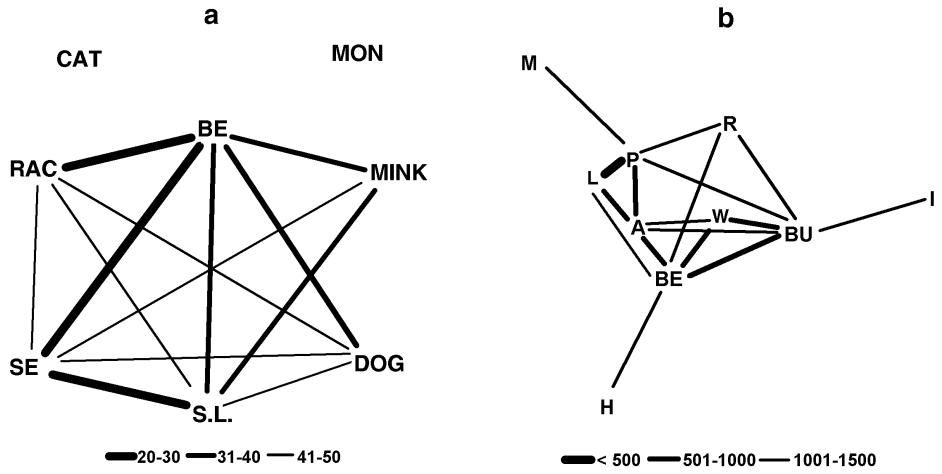


Figure 8.10. Plexus graphs from the matrix of immunological distances of mammals (**a**) and from the distance matrix A5 of European cities (**b**). The distance values are arbitrarily categorized in both diagrams, but in practice the edges are usually colored according to formal significance levels (when available, as is for χ^2 and correlation).

by Whittaker (1987), Moskát (1991) and Dale (2000). This final note leads us directly to the evaluation and comparison of results, the last big topic to be discussed in this book (see the section on the comparison of results of the different type in the next chapter).