6
How to Obtain Proximities

Proximities are either collected by directly judging the (dis-)similarity of pairs of objects, or they are derived from score or attribute vectors associated with each of these objects. Direct proximities typically result from similarity ratings on object pairs, from rankings, or from card-sorting tasks. Another method, called the anchor stimulus method, leads to conditional proximities that have a restricted comparability and require special MDS procedures. Derived proximities are, in practice, most often correlations of item scores over individuals. Because there is so much work involved in building a complete proximity matrix, it is important to know about the performance of incomplete proximity matrices (with missing data) in MDS. It turns out that MDS is quite robust against randomly distributed missing data. MDS is also robust when used with coarse proximities, for example, dichotomous proximities.

6.1 Types of Proximities

MDS procedures assume that proximities are given. How one collects these proximities is a problem that is largely external to the MDS procedures discussed in this book.[1] However, because proximities are obviously needed,

[1]Some authors (e.g., Müller, 1984) approach MDS axiomatically. They formulate relational systems that, if satisfied, guarantee the existence of certain forms of MDS representations. Ideally, these axioms can be directly assessed empirically by asking the

and because the way these proximities are generated has implications for the choice of an MDS model, we devote some space to this topic.

In the previous chapters, we encountered different forms of proximities. For example, the proximities in Table 2.1 were distances generated by direct measurement on an atlas. In all other cases, the proximities were but distance *estimates* related to distances by some MDS model. The color similarity data in Table 4.1 were collected by averaging similarity ratings (0 = no similarity, ..., 4 = identical) for all different pairs of colors over all subjects. The Morse code proximities in Table 4.2 were obtained by computing the relative frequencies of "same" and "different" judgments for all pairs of Morse codes over different subjects. The data in Table 4.4 that indicate the similarity of facial expressions are based on scaling dissimilarity assessments for all pairs of faces over all subjects by the method of successive intervals.

These examples all involve some form of *direct* (dis-)similarity assessment for its object pairs, be it ratings on a scale from "no similarity" to "identical", judgments of "same" or "different", or orderings of object pairs on a similarity scale.

In practice, such direct approaches are rather atypical. Proximities usually are not based on direct similarity judgments, but rather are indices *derived* from other information. The most prominent ones are correlation coefficients, such as the product-moment correlations in Table 5.1 that assess the similarity of intelligence test items.

6.2 Collecting Direct Proximities

Direct proximities arise from directly assessing a binary relation of similarity or dissimilarity among the objects.[2] There are many possible ways to collect such data. The most obvious method is to ask respondents for a similarity judgment.

Some Varieties of Collecting Direct Proximities

The most popular method for collecting direct proximities is to *rate* the object pairs with respect to their overall similarity or dissimilarity. Krantz

---

subjects for simple judgments, such as partitioning every subset of at least three stimuli into two groups of relatively similar stimuli. In such an approach, the data collection is intimately related to the axiomatization of the MDS model.

[2]In order to keep the discussion uncluttered, we skip the case of dominance data in this section. Dominance data assess which object in a pair of objects dominates the other one in some sense, such as, for example, preference. They are treated later when discussing unfolding models in Part III.

and Tversky (1975), for example, wanted proximities for pairs of rectangles. They read the following instruction to their subjects (p. 14).

> *In this experiment we will show you pairs of rectangles and we'll ask you to mark an X in the appropriate cell on the scale from 1 to 20 [answer booklet was before subject] according to the degree of dissimilarity between rectangles.*
>
> *For example: if the rectangles are almost identical, that is, the dissimilarity between them is very small, mark X in a low-numbered cell. In the same fashion, for all intermediate levels of dissimilarity between the rectangles, mark X in an intermediate-numbered cell.*
>
> *We are interested in your subjective impression of degree of dissimilarity. Different people are likely to have different impressions. Hence, there are no correct or incorrect answers. Simply look at the rectangles for a short time, and mark X in the cell whose number appears to correspond to the degree of dissimilarity between the rectangles.*

This method of gathering proximities is called *pairwise comparison*. The subject rates every pair of objects on a dissimilarity scale.

Instead of ratings, market researchers often use some method of *ranking* the object pairs in terms of their overall similarity. For that purpose, each object pair is typically presented on a card. The subject is then asked to sort these cards so that the most similar object pair is on top of the card stack and the most dissimilar one at the bottom.

A complete ranking often may be too demanding a task or too time-consuming. Indeed, respondents often have difficulty ranking nonextreme objects. Thus, the "intermediate" ranks may be unreliable. It therefore makes sense to soften the ranking procedure as follows. The respondent is asked first to sort the cards into two stacks (not necessarily of equal size) one containing "similar" pairs and the other containing "dissimilar" pairs. For each stack, this sorting can be repeated until the respondent feels that it becomes too difficult to further partition a given stack into similar and dissimilar objects. The stack with the most similar objects is then scored as 1, the stack containing the next most similar objects as 2, and so on. The object pairs are given as proximities the score of the stack to which they belong. This usually leads to a weak rank-order (i.e., one containing ties), but that is no problem for MDS.

In *Q-sort* techniques (Stephenson, 1953), the respondents are asked to sort the cards with the object pairs into the categories of a scale that ranges, for example, from "very similar" to "not similar at all". The sorting must be done so that the stack on each scale category contains a preassigned number of cards. Typically, these numbers are chosen such that the card stacks are approximately normally distributed over the scale, with few cards

at the extremes and many cards in the middle. Computer programs exist that support this type of data collection.

Free sorting, in contrast, imposes a minimum number of constraints onto the respondents. They are simply asked to sort the cards onto different stacks so that cards showing object pairs that appear similar in some sense are in the same stack. The number of stacks is not specified. It can range from just one stack for all cards to the case where each stack contains only one card. To pairs of objects that are on the same stack, we assign a dissimilarity of 0, and for pairs of objects on different stacks, a 1 (see below, Section 6.5 on co-occurrence data). The advantage of this method is that the subject's task is not demanding, even for a large number of objects, and subjects report to enjoy the task.

Another technique for collecting direct proximities is the *anchor stimulus method*. Given n objects, one object is picked as a fixed comparison A, and the subject is asked to judge the similarity of all other $n-1$ objects to A. Each and every object serves, in turn, as an anchor. This leads to n sets with $n-1$ proximities each. The proximities resulting from the anchor stimulus method are *conditional* ones. Two proximities resulting from the anchor stimulus method have a meaningful relation only if they have the anchor stimulus as a common element. Thus, for example, the proximity for A and X and the proximity for A and Y can be compared because they share the anchor stimulus A. However, comparing the proximity for A and X with the proximity for B and Y (with A and B anchor stimuli) does not make sense, because the anchor stimuli are different. Hence, such data require particular MDS methods, with weaker loss functions that only assess, point after point, how well the distances of each anchor point to all other points represent the respective proximities. The relations of distance pairs that involve four different points are irrelevant.

Conditional data have the advantage that less data have to be ranked at the same time. Instead of ranking $\binom{n}{2}$ different pairs of objects, the anchor method only needs to rank $n-1$ pairs of objects at one time. The task of conditional ranking relative to fixed anchors is easier and yields more reliable data. These data, however, require more judgments altogether and are less comparable.

A systematic comparison among several methods for collecting direct proximities was done by Bijmolt and Wedel (1995). They found that free sorting and pairwise comparisons rate positively with respondents whereas collecting conditional data was considered to be boring and fatiguing. In terms of the data quality and the quality of the MDS solution, pairwise comparisons ranked best followed by free sorting.

On Ordering Object Pairs for Collecting Direct Proximities

The perceived similarity of two objects may depend on the order in which they are presented. For example, we note in Table 4.2 that the Morse code

signal for I is more frequently confused with a subsequent A (64%) than A is with a subsequent I (46%). Tversky (1977) gives another example: it seems likely that North Korea is assessed as similar to Red China, but unlikely that someone feels that Red China is similar to North Korea. Other order effects may arise if certain objects are presented relatively often in a given section of the data collection. For example, if the Morse code for A appears in the first 20 comparisons, it is most likely to have some anchoring effect.

Position effects can be reduced by randomly picking which of the objects of a pair will be in first position. This method avoids that a given object is always first or second in those pairs where it appears. *Timing effects* can be balanced by picking a random order for the object pairs.

An alternative approach is to balance position and timing effects by explicit planning. Ross (1934) developed a method for that purpose. It generally should be superior to the random method if the number of objects is relatively small. A computer program for Ross ordering was written by Cohen and Davison (1973).

Planned Incomplete Data Designs

One of the more obvious obstacles for doing an MDS analysis is that one needs many proximities, which are expensive to collect. The cheapest way to reduce the labor involved in data collection is to replace data by assumptions. Two assumptions are typical in MDS applications. First, it is taken for granted that the proximities are essentially symmetric. This obviates the need to collect both p_{ij} and p_{ji}. Second, the proximity of an object to itself, p_{ii}, is also not assessed empirically, because it seems even more justified to consider this information trivial: the dissimilarity of an object to itself is assumed to be essentially zero. For an MDS program, it is sufficient to have the proximities for one half-matrix.

However, even with a half-matrix, one needs to assess $\binom{n}{2} = n(n-1)/2$ proximities. The quantity $\binom{n}{2}$ grows rapidly with n. For example, for $n = 10$ one needs to collect 45 proximities, whereas for $n = 20$ one needs 190 proximities. Few subjects would be willing or able to rank 190 pairs of objects with respect to their global similarity. Hence, the need for incomplete data collection becomes obvious. Some structured incomplete designs are displayed in Table 6.1 (after Spence, 1983).

How should one plan an incomplete data design? A good solution is to *randomly* eliminate a certain proportion of cells in the proximity matrix and define them as missing data. Spence and Domoney (1974) studied this question in detail. They computed the distances in a given MDS space with dimensionality t, and then took these distances as input to MDS in order to see how well they would be reconstructed by MDS in t dimensions under a variety of conditions. One of these conditions was to add random error to the distances. Another one was to define some of the proximities as

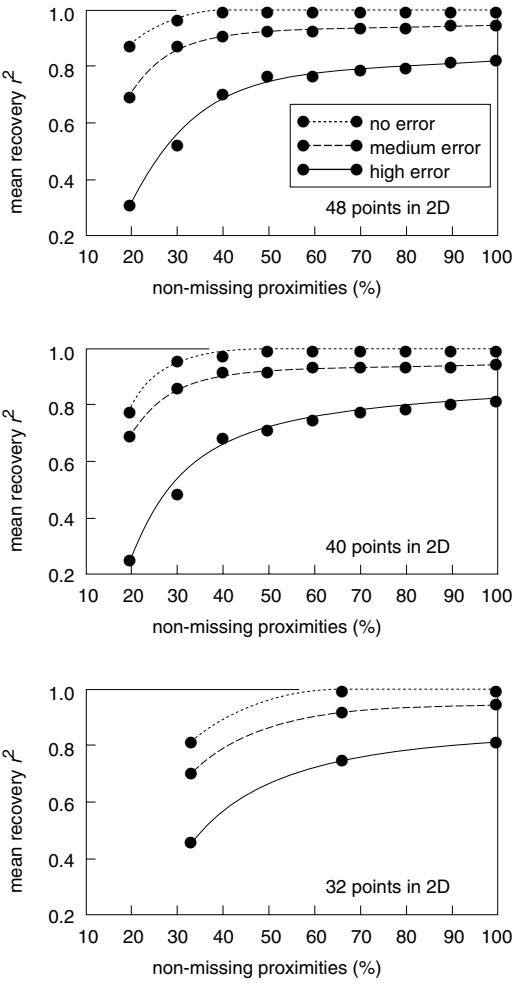


FIGURE 6.1. Recovery of MDS distances (Y-axis) among 48, 40, and 32 points, respectively, under different error levels (upper curves = no error, lower curves = high error) and percentages of nonmissing data (X-axis) (after Spence & Domoney, 1974).

TABLE 6.1. Examples of some incomplete designs (after Spence, 1983). A 0 indicates absence of a proximity, a 1 presence of the proximity.

(a) Cyclic design

	1	2	3	4	5	6	7	8
1	-							
2	1	-						
3	0	1	-					
4	0	0	1	-				
5	1	0	0	1	-			
6	0	1	0	0	1	-		
7	1	0	1	0	0	1	-	
8	1	1	0	1	0	0	1	-

(b) Random design

	1	2	3	4	5	6	7	8
1	-							
2	1	-						
3	0	1	-					
4	0	1	0	-				
5	1	0	1	0	-			
6	1	1	0	0	1	-		
7	0	0	1	0	1	0	-	
8	1	0	1	1	0	0	1	-

(c) Block design

	1	2	3	4	5	6	7	8
1	-							
2	1	-						
3	1	1	-					
4	1	1	1	-				
5	0	0	1	1	-			
6	0	0	1	1	1	-		
7	0	0	0	1	1	1	-	
8	0	0	0	1	1	1	1	-

missing data. It was found that the MDS-reconstructed distances remain highly correlated ($r^2 = .95$) with the original distances if one-third of the proximities are randomly eliminated (i.e., defined as missing data) and the error component in the proximities is about 15%. For high error (30%), r^2 is still .75, which compares well with $r^2 = .83$ for complete data. A significantly greater loss is incurred if two-thirds of the data are missing. However, if the error level is low, excellent recovery is possible even with 80% (!) missing data, given that we scale in the "true" dimensionality t, and given that the number of points is high relative to the dimensionality of the MDS space (see Figure 6.1, upper panels, curves for "no" and "medium" error).

Graef and Spence (1979) showed, moreover, that MDS configurations are poorly recovered if the proximities for the largest distances are missing, whereas missing data for intermediate or short distances are not that crucial. Hence, a missing data design could be improved by making sure that missing data are rare among the proximities for the most dissimilar objects.

These simulation studies show that robust MDS is possible even with many missing data. The user is well advised, nevertheless, to make sure that the missing cells do not form clusters in the proximity matrix.

One should keep in mind, however, that the above simulation results rest on some conditions (many points, reasonable error in the data, known "true" dimensionality, etc.) which are, in practice, often rather difficult to assess. It may be easiest to determine the error level of the data. For direct proximities, it could be estimated by replicating the data collection for some subjects; for correlations, one could consider statistical confidence intervals. Other conditions, however, are less easily diagnosed. For example, the very notion of "true" dimensionality remains obscure in most applications, except in rare cases such as, for example, perceptual studies in a psychophysical context (see Chapter 17). This makes it impossible to come up with a simple answer to the question of how many missing data can be accommodated in MDS.

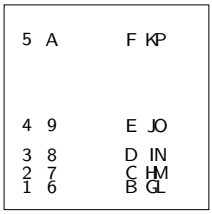


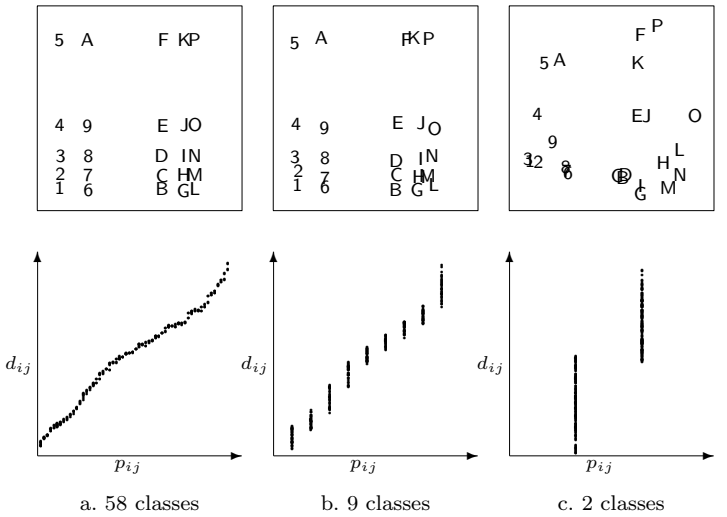
FIGURE 6.2. Synthetic configuration (after Green & Wind, 1973).

Collecting Coarse Data

Another possibility to make the task of collecting proximities simpler in case of direct proximities is to ask the respondents for simpler judgments. One extreme case is the "same" and "different" judgments on the Morse codes (see Chapter 4). Rothkopf (1957) aggregated these judgments over respondents and then analyzed the confusion probabilities as proximities. But is aggregation necessary? Would it make sense to do an MDS on the same–different data of a single individual? At first sight, such data seem "too coarse," but are they?

Green and Wind (1973) report a simulation study that throws some light on this question. They measure the distances of a 2D MDS configuration consisting of 25 points (Figure 6.2). These distances are classified into a small set of intervals. The same ranking number is substituted for all distances within the same interval. The resulting "degraded" distances are taken as proximities in MDS. Using the primary approach to ties (see Sections 3.1, p. 40, and 9.4), it is found that degrading distances into nine ranking numbers still allows one to recover almost perfectly the original configuration (Figure 6.3, Panel b). Even under the most extreme degradation, where the distances are mapped into only two ranking numbers, the original configuration is roughly recovered. One can conclude, therefore, that data that only represent the true distances in terms of distance groupings or blocks can be sufficient for recovering an underlying MDS configuration.

Of course, the granularity of the data may also be too fine in the sense that the data are not reliable to the same extent. For example, in the case of the above 21-point similarity scale employed by Krantz and Tversky (1975), one may well question that the respondents are able to make such fine-grained distinctions. If they are not, then they may not use all of the 21 categories; or if they do, their ratings may not be very reliable. One should not expect that persons are able to reliably distinguish more than 7 ± 2 categories (Miller, 1956). Confronting the individual with a 21-point similarity scale may then actually make his or her task unreasonably difficult.



a. 58 classes b. 9 classes c. 2 classes

FIGURE 6.3. Ordinal MDS representations of distances derived from Fig. 6.2, with Shepard diagrams (after Green & Wind, 1973): (a) uses distances as proximities; (b) uses distances degraded to nine values; (c) uses distances degraded to two values.

Again, there is no rule by which the issue of an optimal granularity could be decided in general. The issue lies outside of MDS, but it is comforting to know that even coarse data allow one to do an MDS analysis. What is important is the reliability of the data.

6.3 Deriving Proximities by Aggregating over Other Measures

Derived proximities are typically correlations or distances computed for a pair of variables, X and Y. A common way to organize the various coefficients available in this context is to consider the scale levels of X and Y. However, in the following, we do not intend to give an encyclopedic overview, but rather present some of the coefficients found most often in the MDS literature. We also discuss a few of the more exotic cases, because they help us to show some of the considerations involved in choosing a proper proximity measure. The obvious scale-level issues are largely ignored.

Correlations over Individuals

Probably the most common case of derived proximities is the one illustrated by the item intercorrelations in Table 5.1. The correlation between item X and item Y is computed over N individuals; that is,

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{(\sum_{i=1}^{N}(x_i - \bar{x})^2)^{1/2}(\sum_{i=1}^{N}(y_i - \bar{y})^2)^{1/2}},$$

where $\bar{x}$ (resp. $\bar{y}$) is the average over all x_is (resp. y_is). A correlation expresses the extent to which the individuals' responses to two items tend to have a similar pattern of relatively high and low scores.

Correlation coefficients exist for assessing different types of trends. The Pearson correlation measures the extent to which two items are linearly related. Substituting ranks for the raw data yields a rank-linear coefficient, Spearman's ρ. It assesses the monotonic relationship of two items. An alternative that responds more smoothly to small changes of the data is the μ_2 coefficient (Guttman, 1985). It is often used in combination with ordinal MDS (see, e.g., Levy & Guttman, 1975; Elizur et al., 1991; Shye, 1985) because (weak) monotonic coefficients are obviously more consistent with ordinal MDS than linear ones. The formula for μ_2 is

$$\mu_2 = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}(x_i - x_j)(y_i - y_j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}|x_i - x_j||y_i - y_j|}.$$

The relationship of μ_2 to the usual product-moment coefficient r becomes most transparent if we express r as

$$r = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}(x_i - x_j)(y_i - y_j)}{\left(\sum_{i=1}^{N}\sum_{j=1}^{N}(x_i - x_j)^2\right)^{1/2}\left(\sum_{i=1}^{N}\sum_{j=1}^{N}(y_i - y_j)^2\right)^{1/2}}$$

(Daniels, 1944). One notes that the denominator of r is never smaller than the denominator of μ_2, which follows from the Cauchy–Schwarz inequality for nonnegative arguments: $\sum_k a_k b_k \leq (\sum_k a_k^2)^{1/2}(\sum_k b_k^2)^{1/2}$. Hence, $|\mu_2| \geq |r|$. One obtains $\mu_2 = r$ exactly if X and Y are linearly related (Staufenbiel, 1987).

Proximities from Attribute Profiles

Correlations typically are computed over individuals; that is, the data in the typical person × variables data matrix are correlated over the rows to yield the intercorrelations of the variables.

Assume now that we want to assess the perceived similarities among a number of cars. One way of doing this is to ask N respondents to assess each of the cars with respect to, say, its attractiveness. Proximities could then be

computed by correlating over the respondents' scores. One notes, however, that this approach completely hinges on the criterion of attractiveness. We may get more meaningful proximities if we do not rely that much on just one criterion but rather on a large selection of attributes on which cars are differentiated. Thus, we could ask the respondents to scale each car with respect to several criteria such as performance, economy, luxury, and so on. (In order to avoid redundancy, one could first factor-analyze these attributes and replace them by supposedly independent criteria or factors.) This would yield a person × cars × attributes matrix. The similarities of cars would then be derived as some function of how similar these cars are over the various attributes.

One possibility is to correlate the attribute profiles of the cars, either for each person in turn (yielding one similarity matrix per person) or over all attributes and all persons (yielding but one global similarity matrix).

An alternative is to measure dissimilarity by computing distances among attribute vectors. Assume, for example, that $\mathbf{X}$ is a cars × attributes matrix that contains average attribute assessments of N persons for each car on m attributes. For example, an element of $\mathbf{X}$ could be the average of the subjective prestige ratings that N persons gave car i. A "simple" distance of any two cars, i and j, in this m-dimensional attribute space is the *city-block distance*,

$$d_{ij}^{(1)}(\mathbf{X}) = \sum_{a=1}^{m} |x_{ia} - x_{ja}|,$$

where i and j are two objects of interest, and x_{ia} and x_{ja} are the scores of these objects on attribute a. Other distances (e.g., the Euclidean distance) are also conceivable but probably less attractive for deriving proximities because they all involve some kind of weighting of the intraattribute differences $x_{ia} - x_{ja}$. For example, in the Euclidean distance,

$$d_{ij}^{(2)}(\mathbf{X}) = \left(\sum_{a=1}^{m} (x_{ia} - x_{ja})^2 \right)^{1/2},$$

the difference terms $x_{ia} - x_{ja}$ are weighted quadratically into the distance function.

An overview of popular proximity measures is given in Table 6.2. To see how the coefficients are related to the attributes, Figure 6.4 shows various isoproximity contours for the case where point x_j is fixed at position $(1, 2)$ and point x_i takes on different positions in the attribute space. The contour lines show the sets of positions where x_i has the same proximity to x_j. In the case of the Euclidean distance, these contours correspond to the usual notion of circles. In the case of the city-block distance, these circles look unfamiliar (see Section 17.2 for more details). On the other hand, the composition rule by which the differences of i and j are aggregated into

TABLE 6.2. Summary of measures of proximities derived from attribute data. The symbol δ_{ij} denotes a dissimilarity and s_{ij} a similarity.

Measure		Formula
P1	Euclidean distance	$\delta_{ij} = \left(\sum_{a=1}^{m}(x_{ia} - x_{ja})^2\right)^{1/2}$
P2	City-block distance	$\delta_{ij} = \sum_{a=1}^{m} \lvert x_{ia} - x_{ja}\rvert$
P3	Dominance distance	$\delta_{ij} = \max_{a=1}^{m} \lvert x_{ia} - x_{ja}\rvert$
P4	Minkowski distance	$\delta_{ij} = \left(\sum_{a=1}^{m}(x_{ia} - x_{ja})^p\right)^{1/p}$ with $p \geq 1$
P5	Canberra distance	$\delta_{ij} = \sum_{a=1}^{m} \frac{\lvert x_{ia} - x_{ja}\rvert}{\lvert x_{ia} + x_{ja}\rvert}$
P6	Bray–Curtis distance	$\delta_{ij} = \frac{\sum_{a=1}^{m} \lvert x_{ia} - x_{ja}\rvert}{\sum_{a=1}^{m}(x_{ia} + x_{ja})}$
P7	Chord distance	$\delta_{ij} = \left(\sum_{a=1}^{m}(x_{ia}^{1/2} - x_{ja}^{1/2})\right)^{1/2}$
P8	Angular separation, congruence coefficient	$s_{ij} = \frac{\sum_{a=1}^{m} x_{ia}x_{ja}}{\left(\sum_{a=1}^{m} x_{ia}^2\right)^{1/2}\left(\sum_{a=1}^{m} x_{ja}^2\right)^{1/2}}$
P9	Correlation	$s_{ij} = \frac{\sum_{a=1}^{m}(x_{ia} - \bar{x}_i)(x_{ja} - \bar{x}_j)}{\left(\sum_{a=1}^{m}(x_{ia} - \bar{x}_i)^2\right)^{1/2}\left(\sum_{a=1}^{m}(x_{ja} - \bar{x}_j)^2\right)^{1/2}}$
P10	Monotonicity coefficient μ_2	$s_{ij} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}(x_i - x_j)(y_i - y_j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\lvert x_i - x_j\rvert\lvert y_i - y_j\rvert}$

the overall distance is extremely simple: the distance is just the sum of the intradimensional differences. The dominance distance, in contrast, is completely determined by just one intradimensional difference of i and j, the largest one. Note that P1 to P3 are special cases of the Minkowski distance P4: $p = 1$ gives the city-block distance P2, $p = 2$ the Euclidean distance P1, and $p = \infty$ the dominance distance P3. The distances P1 to P4 combine dimensional differences directly. Consequently, if the dimensions are attributes measured on different scales, the attributes with the largest variance will dominate the distance measure. Therefore, it is usually better to standardize the attributes so that their variances become equal by converting each attribute to z-scores. Alternatively, each attribute can be divided by another measure for dispersion such as the range (the difference of maximum and minimum).

The proximity measures P5 to P10 all have some provision for controlling the dispersion either for each variable separately or for all variables simultaneously. The Canberra distance corrects the absolute difference along each

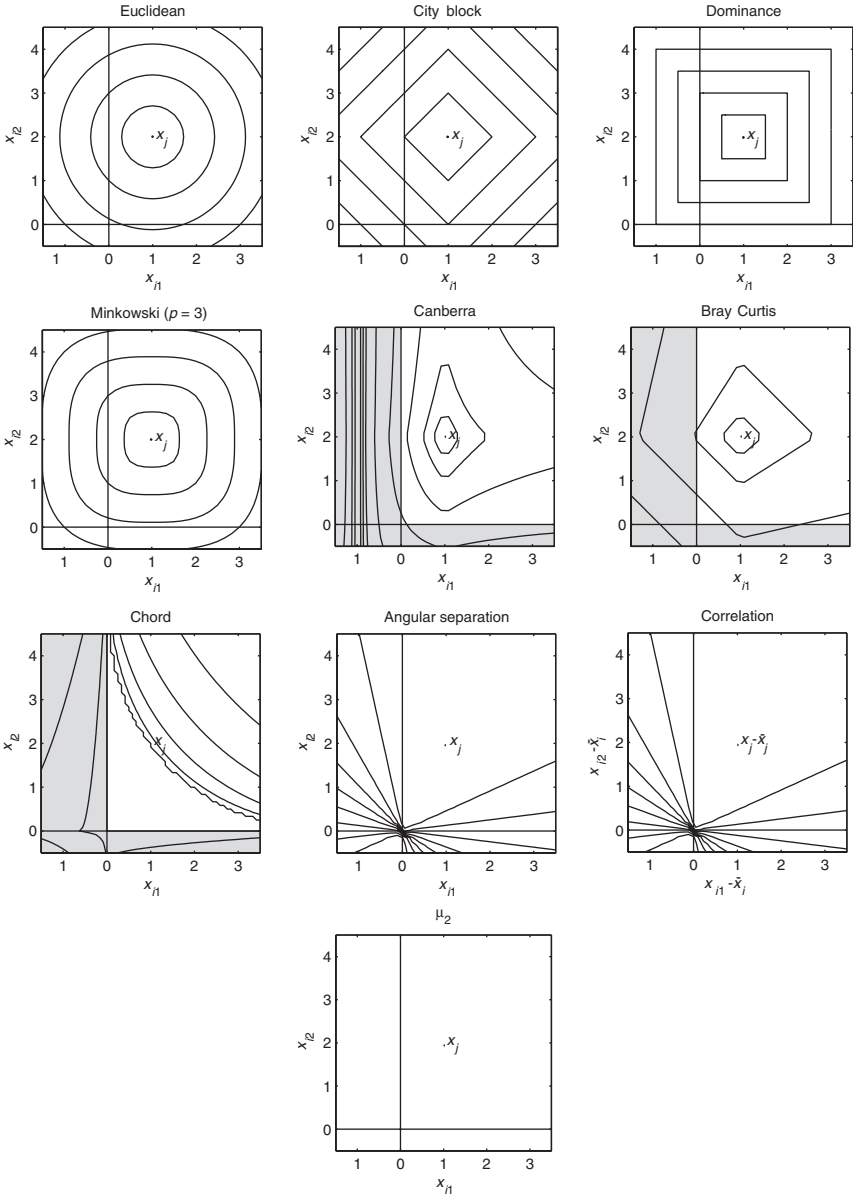


FIGURE 6.4. Contour plots for the different proximity measures defined in Table 6.2, setting $x_j = (1, 2)$. Contour lines close to x_j have low values, whereas further away they have higher values. For the contour lines of the Minkowski distance, the value $p = 3$ was used. Note that μ_2 has no contour lines in this grossly simplified example, because all values are exactly one. The grey areas correspond to negative x_{i1} or x_{i2} which are usually excluded for these measures.

dimension for the size of the coordinates along the axis. In addition, if negative values of x_{ia} are allowed, then δ_{ij} reaches an asymptote of infinity, in Figure 6.4 at $x_{i1} = -1$. Therefore, the Canberra distance is best used when all x_{ia} are positive. The Bray–Curtis distance is often used in ecology and corrects the sum of absolute differences along the axes by the sum of all coordinates over which the differences are taken. Again, this measure seems most useful for nonnegative x_{ia}. In this case, the Bray–Curtis distance corrects large absolute differences when the coordinates are large, too. The chord distance requires positive x_{ia}. Usually, x_{ia} equals the frequency, so that it is positive by nature. Note that for drawing the contour lines in Figure 6.4 for the chord distance, the absolute values of x_{ia} were used. The angular separation is a similarity index between -1 and 1 because it computes the cosine of the angle between the lines from the origin to x_i and the origin to x_j. The contour lines for the correlation are exactly the same as for the angular separation because we changed the axes to $x_{ia} - \bar{x}_i$. Note that both the correlation and μ_2 are best used when the number of dimensions m is reasonably large, certainly larger than in the simplified case of $m = 2$ in Figure 6.4. For μ_2 this simplification leads to $\mu_2 = 1$ for all x_{ia} which explains why there are no contour lines for μ_2. Thus, μ_2 is only meaningful if $m \geq 3$.

Another type of distance function often used in the literature is to count the number of *common elements* in the data profiles and subtract this sum from the total number of attributes on which observations were made. This distance function could be employed, for example, where attributes are coded as either present or absent. An example from archaeology is data on sites where certain artifacts such as pottery, jewelry, bones, and the like, are or are not found (e.g., Kendall, 1971). Sites are considered similar if they share many artifacts.

Restle (1959) suggested this distance function in order to model the perception of similarity: conceiving stimuli X and Y in terms of "feature" sets (i.e., as collections of the things associated with them), we have the distance $d_{XY} = m(X \cup Y) - m(X \cap Y)$, where m is a measure function.[3] Hence, the dissimilarity of X and Y, d_{XY}, is the number of their noncommon features, $m(Y - X) + m(X - Y)$.

When collecting object $\times$ attribute data sets in real life, some attributes may be binary; others may be numerical. The general similarity measure of Gower (1971) is particularly suited for this situation. Let s_{ija} be the similarity between objects i and j on variable a. For binary attributes, we assume that only values $x_{ia} = 0$ and $x_{ia} = 1$ occur. In this case, $s_{ija} = 1$ if x_{ia} and x_{ja} fall in the same category and $s_{ija} = 0$ if they do not. If

[3] A simple measure function is, for example, the number of elements in the set. $X \cup Y$ is the union of X and Y; $X \cap Y$ is the intersection of X and Y; $X - Y$ is the set consisting of the elements of X that are not elements of Y.

the attribute is numerical, then we compute $s_{ija} = 1 - |x_{ia} - x_{ja}|/r_k$ with r_k being the range of attribute a. This definition ensures again that $0 \leq s_{ija} \leq 1$ for all combinations of i, j, and a. The general similarity measure can be defined by

$$s_{ij} = \frac{\sum_a w_{ija} s_{ija}}{\sum_a w_{ija}},$$

where the w_{ija} are given nonnegative weights. Usually w_{ija} is set to one for all i, j, and a. However, if either x_{ia} or x_{ja} is missing (or both), then w_{ija} should be set to zero so that the missing values do not influence the similarity. Here, too, $0 \leq s_{ij} \leq 1$ so that dissimilarities can be obtained by taking $1 - s_{ij}$. However, Gower (1971) suggests to use $(1 - s_{ij})^{1/2}$ as it can be shown that these values can be perfectly represented in a Euclidean space of high dimensionality.

6.4 Proximities from Converting Other Measures

Derived proximities are not always computed by aggregating over individuals or from aggregating over attribute vectors associated with the objects of interest. They can also be generated by appropriate conversion of given scale values for the objects. The conversion is arrived at by theoretical considerations.

Consider the following case. Glushko (1975) was interested in the "goodness" of patterns. He constructed a set of different dot patterns and printed each possible pair on a separate card. Twenty subjects were then asked to indicate which pattern in each pair was the "better" one. The pattern judged better in a pair received a score of 1, the other one a 0. These scores were summed over the subjects, and a dissimilarity measure was constructed on the basis of the following logic. "Since dissimilar goodness between two patterns is implied by frequent choice of either one over the other, the absolute value of the difference between the observed and the expected frequency of a goodness preference represents the dissimilarity of the pattern of goodness of the two patterns ..." (Glushko, 1975, p. 159). Because there were 20 subjects, the expected (random) preference value is 10 for each pair. Hence, proximities were derived by subtracting 10 from each summation score and taking its absolute value.

A similar conversion is the following. Thurstone (1927), Coombs (1967), and Borg (1988) asked N students to indicate in a pair-comparison design which of two offenses (such as murder, arson, or theft) was more "serious." Scoring the more serious one as 1 and the other one as 0, adding these scores over individuals, and dividing by N, one obtains a matrix of dominance probabilities (P_{ij}). These data typically are scaled by Thurstone's Law of Comparative Judgment model, which relates the P_{ij}s to scale values by a cumulative normal density function. However, one can also convert

the probabilities into dissimilarities δ_{ij} and then use ordinal MDS. [Ordinal MDS does not assume a particular (monotonic) model function and, thus, leaves it to the data to exhibit the exact shape of the transformation function.] The conversion formula is $\delta_{ij} = |P_{ij} - 0.5|$.

Tobler and Wineburg (1971) report another interesting proximity, a measure of social interaction between towns or "places" called the *gravity model*: $I_{ij} = kP_iP_j/d_{ij}^2$, where "$I_{ij}$ is the interaction between places i and j; k is a constant, depending on the phenomena; P_i is the population of i; P_j is the population of j; and d_{ij} is the distance between places i and j. Distance may be in hours, dollars, or kilometers; populations may be in income, numbers of people, numbers of telephones, and so on; and the interaction may be in numbers of letters exchanged, number of marriages, similarity of artifacts or cultural traits, and so on." (p. 2). With measures for I_{ij}, P_i, and P_j, the gravity model can be used to solve for the distance d_{ij}. Tobler and Wineburg (1971) report an application from archaeology. Cuneiform tables from Assyria were the database. The number of occurrences of a town's name on these tables was taken as P_i, the number of co-occurrences on the tables as a measure of I_{ij}. The resulting distance estimates were taken as input for a 2D ordinal MDS in an effort to find the (largely unknown) geographical map of these towns.

6.5 Proximities from Co-Occurrence Data

An interesting type of proximities is co-occurrence data. Coxon and Jones (1978), for example, studied the categories that people use to classify occupations. Their subjects were asked to sort a set of 32 occupational titles (such as barman, statistician, and actor) into as many or as few groups as they wished. The result of this sorting can be expressed, for each subject, as a 32×32 *incidence matrix*, with an entry of 1 wherever its row and columns entries are sorted into the same group, and 0 elsewhere. The incidence matrix can be considered a proximity matrix of dichotomous (same–different) data.[4]

Are such co-occurrence data direct proximities? The answer depends on how one wants to define "direct". In the above study on occupation titles, the criterion of similarity should have been obvious to the respondents. Hence, by sorting the occupation titles into groups, they were directly ex-

[4]Burton (1975) further suggests some forms of weighting such as replacing 1 by the number of objects in the category to which a given object pair belongs, or by replacing 1 by the inverse of this number. The former is supposed to emphasize gross discrimination, the latter fine discrimination. Such weightings of global and local discriminations are, however, better introduced as part of the MDS modeling criteria, rather than building them into the data.

pressing their notions of pairwise similarity relations for these stimuli. But consider another case.

England and Ruiz-Quintanilla (1994) asked respondents to check those characteristics in a list that would define work for them. The characteristics were "if it is not pleasant", "if it is physically strenuous", "if you have to do it", and so on. The co-occurrences of these characteristics were defined as the characteristics' proximities. Its seems that this definition is more an interpretation of the researcher, because the respondents never directly assessed the similarity of the characteristics in the context of work, but their relevance with respect to the notion of work. Hence, these proximities seem somewhat more derived than the former ones, which shows that the direct-derived distinction denotes more a continuum than a dichotomy.

Studies that use co-occurrence data typically aggregate incidence matrices over individuals. The most natural way to do this is simply to add these matrices so that the aggregate proximity matrix contains in its cells the frequencies with which two objects were sorted into the same group.

However, it is well worth the effort to consider whether it would be better if these raw frequencies were normed. Let X and Y be two items of interest. An item X can be empirically present or absent, denoted as $X = 1$ and $X = 0$, respectively. With X and Y, there are four possible present–absent combinations. Let $z = f(X, Y)$ be the frequency of an event (X, Y). In particular, let $a = f(1, 1)$ be the frequency of the event where both X and Y are present. Similarly, $b = f(1, 0)$, $c = f(0, 1)$, and $d = f(0, 0)$ (see also Table 6.3). Gower (1985) distinguishes a variety of possible similarity coefficients, all of which vary between 0 and 1. One possibility is

$$s_2 = a/(a + b + c + d),$$

the frequency of events where both X and Y occur relative to the total frequency of all present–absent combinations of X and Y. Another possibility is

$$s_3 = a/(a + b + c),$$

the proportion of events where both X and Y occur, given at least one of them occurs (*Jaccard similarity measure*).

To see the consequences of choosing s_2 or s_3, consider the following example. Bilsky, Borg, and Wetzels (1994) studied forms of conflict tactics among family members, ranging from calm debates over throwing things to physical violence inflicting injuries to other persons. A survey asked the respondents to indicate which forms of behavior had occurred among members of their families in the last five years. If co-occurrence of behavior forms is assessed by s_3, MDS yields a one-dimensional solution where the different behavior forms are simply arrayed in terms of their aggressiveness, with a major gap between behaviors that involve shouting, throwing things, and the like, and those that involve any form of physical violence. Using s_2

TABLE 6.3. Types of combinations of two events X and Y, together with their frequencies (cells entries).

	$X=1$	$X=0$	Total
$Y=1$	a	b	$a+b$
$Y=0$	c	d	$c+d$
Total	$a+c$	$b+d$	$a+b+c+d$

coefficients, however, has the effect that the behaviors that involve physical violence drop dramatically in similarity because they are so rare, that is, because d is so great. This essentially wipes out the clear scale obtained for s_3 proximities.

There are amazingly many ways to combine the four frequencies $a, \ldots, d$ into an overall proximity measure for each pair of objects (see, e.g., Gower, 1985; Gower & Legendre, 1986; Cox & Cox, 1994). However, most of these proximities make sense only in highly specific contexts, so that it serves no purpose to discuss all of them here. It may suffice to consider just one further proximity, the *simple matching coefficient*,

$$s_4 = (a+d)/(a+b+c+d),$$

which counts both co-occurrence and co-nonoccurrence as indices of similarity. In the case of the forms of violent behaviors, s_4 would bring up the question of whether rare forms of behavior, in particular, should be considered very similar simply because of their high rate of co-nonoccurrence. More details about many of the possible binary coefficients and their scalability in MDS can be found in Gower and Legendre (1986).

An small overview of the most frequently used co-occurence measures is presented in Table 6.4, together with the range for each of these indexes. It is easy to convert these similarity measures into dissimilarities by computing $\delta_{ij} = 1 - s_k$, for $k = 2, \ldots, 6$.

6.6 Choosing a Particular Proximity

The availability of so many varieties of proximities seems to make life confusing for the user. Which proximity should be chosen? An answer to this question depends on many considerations, but is typically not that difficult.

An important decision criterion is usually the practical feasibility of a particular data collection method. Consider surveys, for example, where respondents are asked by questionnaires about their attitudes towards various political issues. It would be inconceivable to replace the usual item-by-item ratings by a task where the respondent has to compare the $n(n-1)/2$ pairs of items, because this is simply too time consuming. Moreover, it would be difficult to explain to the respondents what exactly they are supposed to

TABLE 6.4. Overview of some popular co-occurrence measures.

Measure			Bounds of s_k
s_2		$s_2 = \frac{a}{a+b+c+d}$	$0 \leq s_2 \leq 1$
s_3	Jacard similarity measure	$s_3 = \frac{a}{a+b+c}$	$0 \leq s_3 \leq 1$
s_4	Simple matching coefficient	$s_4 = \frac{a+d}{a+b+c+d}$	$0 \leq s_4 \leq 1$
s_5	Hamman	$s_5 = \frac{(a+d)-(b+c)}{a+b+c+d}$	$-1 \leq s_5 \leq 1$
s_6	Yule	$s_6 = \frac{ad-bc}{ad+bc}$	$-1 \leq s_5 \leq 1$

do in such a task, that is, in which sense they are supposed to compare the items.

Another case is the proximity of intelligence test items, assessed above in terms of how similarly the testees perform on the items. Here, it remains unclear how direct proximities could be defined at all without changing the research question. Assume that we would ask test psychologists to evaluate directly the global similarity of the test items. Such a question, obviously, studies the perception of test psychologists and not the structure of the test item performance of testees.

Direct proximities are more a task for laboratory studies on perceptual structures than, for example, for survey studies. Most of the examples discussed earlier (e.g., Morse code confusions, color similarities) belong to this category. The card-sorting procedures often used by market researchers is another example.

In the context of such research questions, direct proximities typically are collected to explain how they are generated. If the subjects were asked to first assess the objects of interest on scales invented by the researcher, the proximities would be based on these scales, not on criteria freely chosen by subjects themselves. In the facial expressions study by Engen et al. (1958), the direct proximities were, therefore, collected along with ratings on certain dimensions in order to check whether the structure of the former could be explained by the latter (see Section 4.3).

So, the question of what proximity to choose typically is decided to a large extent by the research question and its context. However, this is more true for direct proximities. If one decides to derive proximities, one has a less substantive foothold for choosing a particular measure.

Deriving proximities requires one to decide, first of all, if one wants a correlation coefficient or a distance measure on the observations on two

variables, X and Y. The former assesses the similarity of X and Y in terms of their "profiles", the latter the (dis-)similarity in terms of their element-by-element differences. That is, if $X = 2 \cdot Y$, for example, then $r_{XY} = 1$, but the distance of X and Y is not zero. On the other hand, if the distance of X and Y is zero, then $r_{XY} = 1$ always.

However, the choice between these measures is not that important in practice. The reason is that if proximities are computed by aggregating over attribute scales, it usually makes sense to first standardize the different attribute scales rather than using raw scores. In this case, Euclidean distances are related to Pearson's r by a monotonic function. This can be seen as follows. Assume that we have two variables, X and Y, that are both standardized so that their means are zero and their sum-of-squares is equal to 1. As a result, $r_{XY} = \sum_i x_i y_i$. Then, the Euclidean distance between X and Y is

$$\begin{aligned} d_{XY} &= \left(\sum_{i=1}^{N} (x_i - y_i)^2 \right)^{1/2} \\ &= \left(\sum_{i=1}^{N} x_i^2 + \sum_{i=1}^{N} y_i^2 - 2 \sum_{i=1}^{N} x_i y_i \right)^{1/2} \\ &= (2 - 2r_{XY})^{1/2} . \end{aligned} \tag{6.1}$$

Hence, when using ordinal MDS, it becomes irrelevant which proximity is used, because both yield (inversely) *equivalent* rank-orders.

City-block distances, moreover, are typically highly correlated with Euclidean distances, so that they, too, are monotonically closely related to r in practice. It is also true that Pearson correlations and monotonic correlations such as ρ or μ_2 are highly correlated if the relationship of the items is not extremely nonlinear. Moreover, the *structural* information contained in a matrix of proximities is very robust against variations in the individual proximity coefficients. For that reason, Pearson rs are often chosen in practice rather than the formally more attractive μ_2s. In summary, then, the user need not worry that much about the particular choice for computing proximities from score vectors: the usual measures, such as r or the Euclidean distance, are most often quite appropriate in an MDS context.

6.7 Exercises

Exercise 6.1 Consider the matrix of dominance probabilities P_{ij} below (Borg, 1988). It shows the relative frequencies with which a group of male students judged that the crime/offense in row i is more serious than the crime/offense in column j. Thurstone (1927) and Coombs (1967) report similar data. They analyze them with the Law-of-Comparative-Judgment

model. This model maps dominance probabilities P_{ij} into scale value differences $x_i - x_j$ by the inverse normal distribution ogive; that is, $N^{-1}(P_{ij}) = x_i - x_j$, where N^{-1} denotes the function that maps probabilities into z-scores.

Item	1	2	3	4	5	6	7	8	9	10
1 Abortion	.50	.65	.32	.30	.42	.12	.20	.36	.45	.49
2 Adultery	.35	.50	.20	.19	.25	.02	.11	.28	.31	.33
3 Arson	.68	.80	.50	.41	.62	.13	.22	.45	.61	.67
4 Assault/battery	.70	.81	.59	.50	.67	.16	.29	.51	.70	.72
5 Burglary	.58	.75	.38	.33	.50	.09	.14	.40	.58	.58
6 Homicide	.88	.98	.87	.84	.91	.50	.59	.74	.87	.90
7 Rape	.80	.89	.78	.71	.86	.41	.50	.63	.83	.83
8 Seduction	.64	.72	.55	.49	.60	.26	.37	.50	.66	.69
9 Theft	.55	.69	.39	.30	.42	.13	.17	.34	.50	.53
10 Receiving stolen goods	.51	.67	.33	.28	.42	.10	.17	.31	.47	.50

(a) Davison (1983) suggests that these data can be modeled by ordinal MDS. In fact, he claims that one can solve for a more general class of models called Fechner models. All Fechner models require that (1) $P_{ij} = 0.5 \mapsto d_{ij} = |x_i - x_j| = 0$ and that (2) $d_{ij} = |x_i - x_j|$ grows strictly monotonically as a function of $\delta_{ij} = |P_{ij} - 0.5|$.] Thurstone's model is but one particular Fechner model that relies on the normal function. Use ordinal MDS to find one-dimensional scales for the crime/offense data sets without relying on any particular monotonic function.

(b) Study the empirical relation of dominance probabilities to the corresponding scale differences (=signed distances) and discuss whether the normal mapping function used in the Law-of-Comparative-Judgment model is empirically supported here.

(c) Repeat the MDS analysis with five different random starting configurations. Compare the five solutions. What does your finding imply for unidimensional scaling?

Exercise 6.2 Consider Table A.1 on page 545 in Appendix A that compares several properties of MDS programs. Drop the rows "max. number of objects", "min. number of objects", and "max. dimensionality" as computer constraints that have little to do with the substance of the different MDS programs described here. Turn the remaining matrix into a 1–0 incidence matrix. Then compute at least three different types of similarity coefficients for the set of MDS programs and discuss your choices. Finally, scale these similarity data in 2D MDS spaces and compare the resulting solutions.

Exercise 6.3 Consider Table 1.5 used in Exercise 1.7.

(a) Derive proximity matrices for the row entries by using (1) monotone correlations, (2) city-block distances, and (3) Euclidean distances.

(b) For each set of proximities, find 2D ordinal and interval MDS solutions.

(c) Compare the solutions: How similar are they? Give reasons for their relative similarities or dissimilarities.

Exercise 6.4 Pick ten countries from at least four different continents. For these countries, derive a proximity matrix by card sorting, where you are the respondent yourself. Discuss which problems you encountered in sorting the cards. Replicate the experiment with a different respondent and compare the outcomes.

Exercise 6.5 Consider the data matrix below. It shows the results of a free sorting experiment reported by Dunn-Rankin (1983, p.47). Fifteen persons clustered 11 words that all begin with the letter "a". The entries in the data matrix are cluster numbers.

Person	A	Ad-mits	Aged	Al-most	Aim-ing	And	As	At	Areas	Army	Away
1	1	2	3	2	4	3	1	1	5	6	6
2	1	2	3	2	2	1	1	1	3	2	2
3	1	2	1	2	2	3	1	1	3	3	3
4	1	2	3	4	4	1	5	6	7	8	8
5	1	2	3	4	4	1	5	6	7	8	8
6	1	2	3	3	4	5	1	6	7	8	8
7	1	2	3	2	2	3	1	1	2	2	2
8	1	2	2	4	5	6	7	7	8	9	9
9	1	2	3	2	4	5	1	6	4	4	4
10	1	2	3	4	5	2	1	1	2	6	6
11	1	2	3	2	4	1	1	1	3	5	5
12	1	2	3	4	2	3	1	1	3	3	3
13	1	2	3	2	4	5	1	1	6	7	5
14	1	2	3	2	4	5	1	1	6	7	7
15	1	2	3	2	2	3	1	1	3	2	3

(a) Do the persons sort the words into the same number of clusters? Which person makes the finest distinctions and which person the coarsest?

(b) Compute a matrix of reasonable proximity indices for the 11 words. Analyze the similarities by MDS.

(c) Compute proximity indices for the 15 persons and analyze the indices by MDS. (Hint: Make a list of all pairs of words. If person x throws word i and word j into the same cluster, assign a proximity score of 1. Else, score 0.)

Exercise 6.6 Merkle (1981) studied the frequencies with which product x is bought together with product y, as measured by the sales registry in a set of German clothing stores. He reports the following co-occurrence data.

Product	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1 Expensive suit	28													
2 Expensive trad. shirt	18	68												
3 Expensive tie	13	17	0											
4 Cheap tie	6	8	0	13										
5 Imported shirt	10	25	10	0	20									
6 Medium-priced shirt	2	23	0	15	3	0								
7 Cheap suit	2	27	6	22	6	13	26							
8 Cheap shirt	3	9	10	25	25	13	26	57						
9 Cheap knitwear	17	46	22	24	5	109	222	275	487					
10 Stylish shirt	10	0	4	8	1	48	146	88	57	109				
11 Colored socks	24	21	3	18	7	281	197	117	178	8	273			
12 Jeans	25	10	23	9	5	43	167	146	46	8	46	110		
13 Modern jacket	1	14	3	33	0	3	12	21	87	42	15	14	508	
14 Modern pants	0	0	0	46	16	0	18	67	12	19	20	24	45	88

(a) Discuss how the values on the main diagonal of this matrix are to be interpreted. Are the data similarities or dissimilarities?

(b) Some products are bought more often than others. Discuss what effects this has if one were to submit these data to an MDS analysis. In which ways would the result be influenced by buying frequencies? Where in the MDS plot would a product move that people tend to buy very often?

(c) Merkle (1981) suggests normalizing these data for their different basic frequencies by using Yule's coefficient of colligation: $Y_{xy} = [\sqrt{ad} - \sqrt{bc}]/[\sqrt{ad} + \sqrt{bc}]$, where a denotes the frequency of all sales that contain both x and y, d is the frequency of sales that contain neither x nor y, b are sales of x but not y, and c are sales of y without x. Compute the Y_{xy} coefficients for co-sales of products 1 through 4.

(d) The coefficient Y_{xy} is not easily interpretable. If, however, one skips the square roots in the formula for Y, another coefficient due to Yule results, called Q (see s_6 in Table 6.4). What does Q assess? How can this be expressed in words?

(e) Assume we wanted to do an ordinal MDS of the normalized data. Would it make much, or any, difference whether we use Y or Q?

(f) Describe alternatives for normalizing the data matrix for different overall sales frequencies of the different products.

(g) Compute MDS solutions for these data, both raw and normalized. Discuss the solutions in terms of what features of these products determine whether they tend to be bought jointly or not.

(h) Make a proposal of how the different values of the main diagonal could be represented graphically in the MDS plots.