5 Characteristics of MDS Models

The old order changeth, yielding place to new
And God fulfils himself in many ways,
Lest one good custom should corrupt the world

TENNYSON (Idylls of the King)

We have used three basic characteristics to define the basic MDS model. These are:

Type of data
Transformation (rescaling) function (or level of measurement)
Form of model.

In this chapter we shall use these characteristics to differentiate and describe other programs in the MDS(X) series. Programs from other sources can be described in precisely the same way. This three-fold characterisation is akin to the new typology of scaling programs developed by Carroll and Arabie (1980) in their definitive review of scaling developments of the past seven years.

We begin by giving a fully-specified description of the basic model and then proceed to a discussion of the various ways in which types of data, transformations and models can be extended. These characteristics will be used to describe the MDS(X) programs for the analysis of 2-way data in the next chapter and for the analysis of 3-way data in Chapter 7.

The basic model: a fuller specification
The basic model has already been defined (section 3.1) as follows:

BASIC NON-METRIC MODEL
The analysis of

(Characteristic)	(Specification)
(DATA)	A square symmetric 2-way table of (dis)similarities
(TRANSFORMATION)	by a monotonic rescaling function
(MODEL)	using a simple Euclidean distance model

Enlarging on this specification we might describe the basic model as in Table 5.1. Each of the aspects appearing here is taken up in the following section.

5.1 Data

By data in this context we simply mean information input to the program. Even for

BASIC NON-METRIC MODEL		EXTENSIONS
Characteristic	Specification	
DATA	The INTERNAL analysis of a 2-WAY SQUARE (Unconditional) SYMMETRIC matrix of (dis)similarities	External 3- and higher way/modes Non-square, rectangular, triadic ...
TRANSFORMATION FUNCTION	By a GLOBALLY MONOTONIC rescaling	Locally Linear, Power, Continuity ...
MODEL	Using a SIMPLE EUCLIDEAN DISTANCE model	Weighted Other Minkowski and non-Minkowski metrics Additive, Subtractive, Multiplicative (scalar product factor) ...

Table 5.1 *Fuller specification of basic MDS Model and extensions*

the basic model, which assumes a square, symmetric (or lower triangular) array of data, the information might either be pairwise similarity ratings obtained directly from subjects, or else indirect measures of co-occurrence, covariance, contingency, association etc., obtained by aggregating over simpler data, as we saw in Chapter 2. The *source* of the data does not concern us; the *form* of the data and its interpretation do.

The distinctions referring to data which are made in Table 5.1 are between:

(i) the way and mode of the data; and
(ii) internal and external analysis

We shall consider each in turn.

5.1.1 The 'way' and 'mode' of data

The 'way' of data is simply the dimensionality of the data array. 'One-way' data would simply consist of measures on a single set of objects, such as *one* individual's set of preference judgments of the loudness of a set of tones, or the frequency of a particular plant species on a set of geographical sites. Since one-way data are never scaled as they stand, they need not be examined in detail here.

Two-way data take the form of a matrix consisting of rows and columns, and relate a pair of entities. To say that a set of data is two-way says only that it may be represented in a single matrix. It does not tell us whether the matrix is square or rectangular, symmetric or asymmetric.

In order to make such distinction, the notion of mode is introduced. In a two-way matrix, the rows and columns may refer to the same set of objects or to distinct

sets. If the rows and columns refer to the same entities—and the matrix is thus necessarily square—then the matrix is said to have *one* mode, the one set of entities represented. If, on the other hand, rows and columns refer to two distinct sets, then the data are said to have two modes. The mode of the data therefore is the number of distinct sets of entities to which it refers.

Normally, of course, two-way, two-mode data form a rectangular matrix with, conventionally, 'data producers' (individuals, groups, locations) as row-elements, and objects (stimuli, attitude statements, symbolic entities) as column-elements; but two instances where this is not the case are apparent, which clarify the usefulness of this way/mode distinction. The first is where the number of row-elements happens to be equal to that of the column-elements, and the second where the row- and column-entities are in fact the same but are considered distinct for the purpose of analysis, e.g. firms considered as producers and consumers, members of a group as rankers and ranked in a sociometric exercise.

The extension to higher ways and modes should be obvious, and these are considered later in Chapter 7.

5.1.1.1 *Asymmetric data*

Of particular interest are those data matrices where the row and column elements happen to refer to the same objects—so the matrix is square—but where the elements δ_{jk} and δ_{kj} are considered distinct. Such data occur as sociometric rankings, occupational-mobility turnover tables, economic input-output tables, migration and communication flows, citations within and between journals, and confusion between pairs of auditory stimuli presented in a left-right and right-left order.

At first sight it may seem perverse to wish to represent such data by what is, after all, by definition a *symmetric* distance model. Several ways have been proposed to deal with this anomaly:

(i) to treat the asymmetry of δ_{jk} and δ_{kj} as 'noise' or chance error and simply symmetrise the data by replacing the corresponding entries by their median, or by the arithmetic or geometric mean. Such a treatment, of course, simply defines the problem out of existence and the resulting symmetrised data matrix can now be analysed by the basic model.

(ii) to treat the asymmetric information as consisting of two distinct components, each of which is capable of being represented separately by the distance model: the 'flow' from j to k and the 'flow' from k to j. This alternative was discussed in section 2.2.3.4 where the index of dissimilarity was used to compare both row percentages (outflow) and column percentages (inflow). Typically, the two resulting matrices of outflow and inflow coefficients are scaled separately by the basic model, and the solutions are then compared (see, for instance, Macdonald 1972, pp. 214–27 and Blau and Duncan 1967, pp. 67–75).

(iii) to treat the asymmetry as arising from the conditional nature of the data, but not as a characteristic needing separate representation. The entries within the same row of the matrix will be treated as being comparable, but information between rows will not.

This interpretation is most relevant where data have been collected by the method of conditional rank orders (see Rao and Katz 1971, p. 470) or, in Coombs' terminology, 'order $(p-1)$ out of p stimuli'. In this method the subject is presented with each stimulus in turn. She is then asked to rank each of the other stimuli in terms of their similarity to the reference stimulus, thus generating what amounts to a set of p I-scales, with each stimulus in turn serving as the 'ideal point'.

In this last instance, the stimuli are represented as a *single* set of points, although the entries δ_{jk} and δ_{kj} will normally be fit by distinct disparity values. (This is the model fit by MINICPA described in section 6.1.2.) Another alternative is to treat the row and column elements as distinct points. Thus each 'stimulus as subject' (rows) and each 'stimulus as object' (columns) will be represented as separate points. This option also treats the data as providing conditional distance information and is identical to the unfolding model described in 5.3.3.1. It is implemented by the MINIRSA program.

(iv) to treat the asymmetry as something extrinsic to the distance information and represent it in some other way. A number of ways have been proposed, including representing asymmetry as contours and as 'jet-stream' directions over a conventional scaling configuration, and are discussed in Gower (1977).

(v) to interpret the data as a graph with each distance represented as a link between two points, allowing the distance $i \rightarrow j$ to be different in length from $j \rightarrow i$ (see 6.1.2).

5.1.2 *Internal* vs *external analysis*

The distinction between internal and external analysis was made in Chapter 4 with regard to the interpretation of configurations. There we noted that in internal analysis it was the original data only that were used in the interpretation while additional information was brought to bear in the external case. Generally speaking, internal analysis, or 'unconstrained' solutions (Carroll and Arabie 1980), uses only the information given to generate the solution, while external analysis ('constrained' solutions) takes one part of the input as fixed and relates the data to that fixed 'external' part.

5.2 Transformations

Whilst the full range of Stevens' levels of measurement may, in principle, be used in scaling, only a small number have in fact been used, and in the MDS context the only ones which concern us directly are the nominal, ordinal, interval and ratio levels. By and large, most data are at the nominal and ordinal level—or researchers with justifiable caution consider their data so to be—whereas most scaling solutions are at the ratio level (e.g. distances) or occasionally at the interval level (e.g. solution scales from conjoint measurement).

The transformation function, rescaling the data into distances, normally matches the level of measurement of the data. For our purposes, the most important transformations are the monotonic (ordinal) and regular (linear or logarithmic) rescaling functions, but we shall also consider the 'continuity' or 'smoothness' transformation, which does not fit easily into the conventional levels of measurement, having affinities with both monotonic and metric scaling.

Although nominal rescaling functions have been developed† they are not widely employed in programs in the MDS(X) series other than SSA(M) and are not considered further.

5.2.1 Monotonic transformations

The monotone relation is best illustrated by the Shepard diagram (Figure 3.11 et seq.), where fitting values (d^0) are joined up to form a jagged monotonic 'curve'. The line segments drawn to join up the fitting value points are simply an aid to visualising the relationship and show that the relationship between the data and distances is ascending (in the case of dissimilarities) or descending (in the case of similarities). But the slope of the segments has no intrinsic meaning whatever since it depends in part on the purely arbitrary ordinal scale of the data.

In most MDS applications the shape of the monotone curve is characteristically very jagged and 'steppy', especially if there are ties in the data and weak monotonicity has been used as a criterion in monotone regression (see section 3.2.3). But users should be alert to signs of smoother regularity in the monotone curve. Two particularly important, and more regular, forms of the monotonic function are the straight line and power function curves (including the exponential and logistic curves). All of these variants of monotone relationships are illustrated in Figure 5.1.

An actual example of a monotone regression function approximating a linear function is seen in the Shepard diagram of Figure 3.2; the co-occurrence data scaled in section 3.6 provide a fair approximation to a (negative) exponential function (Figure 3.14b), and the relation between the rank of a mileage and its recovered distance exemplified in the Scottish mileages data is very well approximated by a logistic function (Figure 3.4). Whenever a more regular function is discerned, the data should then be re-analysed using the appropriate scaling transformation.‡ (Linear and power transformations are permitted in the MRSCAL program).

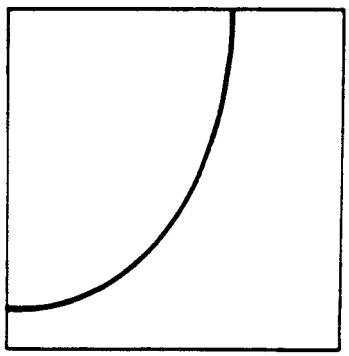
5.2.1.1 Local and global monotonicity

The monotonicity criterion requires that *all* the data should be monotonic with the distances (global monotonicity). On occasion this may be thought too restrictive and monotonicity be only required locally, i.e. around the neighbourhood of each point, hence the term 'local monotonicity'.

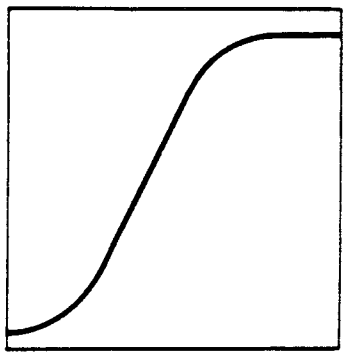
A familiar example of the local monotonicity principle occurs in geographic mapping where 'stereographic projections', which seek to represent the earth's three-dimensional surface as a two-dimensional map, are commonly used. Clearly, this cannot be done without some distortion, and the various geographic projections preserve different aspects of distance. The 'conformal mapping' projection involves a principle very similar to local monotonicity, since smaller distances are accurately represented but larger ones are not. Hence the distortion is

†Nominal rescaling functions are employed in the Multiple Scalogram Analysis option in SSA(M) and in allied programs in the Guttman-Lingoes series (Lingoes et al. 1979, pp. 274–7; Zvulun 1978) and nominal rescaling is permitted as an option in the ALSCAL program (Takane et al. 1977) for analysing 2- and 3-way data.

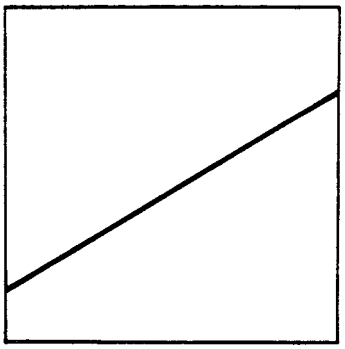
‡Shepard (1974, p. 395 et seq.) describes a number of other approaches to constraining the monotone function to convexity, concavity, smoothness, etc.



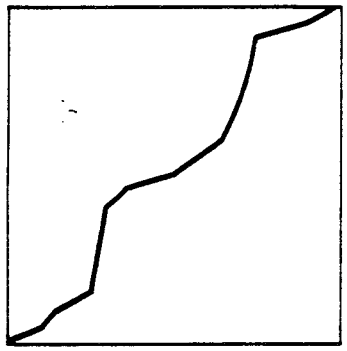
(i) Power or logarithmic
[$y = a^x$; $x = \log_a y$]



(ii) Logistic
[$y = k/(1 + e^{a + bx})$]



(iii) Linear
[$y = a + bx$]



(iv) Irregular monotonic

Figure 5.1 *Monotone curves*

concentrated in the largest distances and this fact needs to be taken into account in reading the map. The same principle applies in MDS solutions derived by use of this criterion.

There are two cases where it is useful to use the local monotonicity criterion, one being when ceiling effects occur in the data producing the familiar C-shape, or horseshoe shape, discussed in section 4.3.2.1. To overcome this effect, the remedy is fairly simple: ignore or down-grade the importance of the largest data dissimilarities. This can be implemented as follows:

(i) by choosing the local monotonicity option in SSA(M); or

(ii) by using a program such as PARAMAP which implements 'continuity' transformation, one of whose features is to act like a local monotonicity constraint (see 5.2.2 below).

Both options have similar and often dramatic effects—in 'unbending' highly non-linear simple structures.

The second use of local monotonicity is to map a high-dimensional solution into a space of lower dimensionality. This procedure is acceptable if local structure is of primary interest and larger distances which will be distorted can be ignored. It

should be noted, however, that Graef and Spence (1979) have shown that the largest distances do most work in producing an MDS solution and they can be critical in the satisfactory recovery of a configuration.

5.2.2 Continuity (smoothness) transformations

In basic non-metric scaling a best overall monotonic fit is sometimes achieved by producing sudden changes in distance values which do not exist in the data values. If we are firmly committed to the assumption that there really is no information in our data other than the order of the dissimilarities, well and good. But if we believe that the data contain more than simple ordinal information then these sudden discontinuous jumps may well distort the local structure, producing high distance values to correspond to very close data values. In this case, it might be better to concentrate on minimising or smoothing out the jumps by making the relationship between the data and the distances of the solution as 'smooth' or 'continuous' as possible, even at the cost of worsening the overall monotonic fit. This can be done by requiring that when two data values are close to each other, then there should be little difference (or variation) in the corresponding distance.

This basic idea of continuity can best be illustrated by a simple example of a one-dimensional solution. Suppose we wish to examine the relationship between the physical loudness of a set of six tones, x_1 to x_6, and their perceived loudness, y_1 to y_6. Our attention will concentrate, as we move up the scale, upon whether perceived differences in loudness change in the same manner as physical differences do.

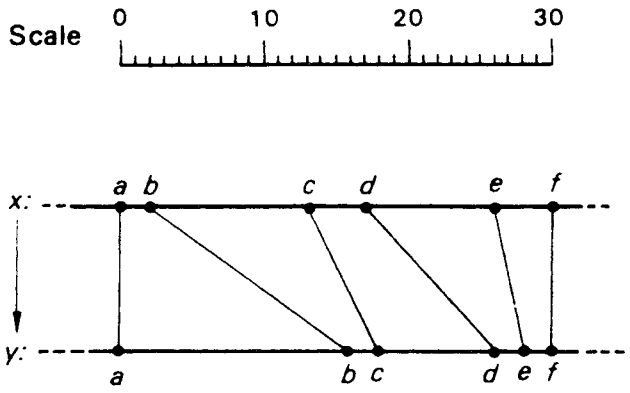
> If we say that the y values seem to change in a 'continuous' manner as we move along the underlying x continuum, we are essentially saying that the change in y as we move from one x value to the next tends to be small compared to the change in y generally associated with larger jumps in x.
>
> (Shepard and Carroll 1966, pp. 579–80)

In Figure 5.2 two examples of such a relationship are given: one where small changes in x are *not* associated with small changes in y, a relatively discontinuous relationship (Figure 5.2a), and another (Figure 5.2b) where small changes in x are associated with small changes in y, a relatively continuous relationship.

The extent to which the relationship between x and y is smooth or continuous can be monitored by a simple index which compares changes in y to changes in x, for each of the adjacent pairs along the scale. This can be done by taking the differences between adjacent values of both x and y and finding the ratio:

$$\frac{\text{Difference in } y}{\text{Difference in } x} = \frac{\Delta y}{\Delta x} = \frac{y_j - y_k}{x_j - x_k}$$

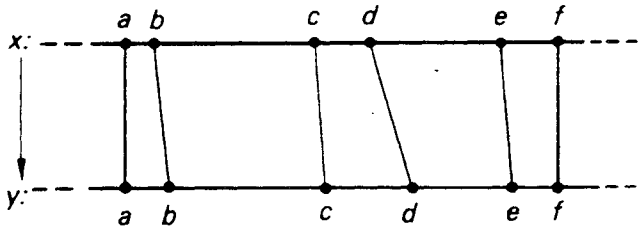
Clearly, when the relationship is smooth or continuous, Δy and Δx will be almost the same and the ratio will be about 1. But if small changes in x produce large changes in y, then the ratio will be correspondingly large. A simple overall measure of discontinuity (DISCONT) is constructed by squaring the ratio (both for computational simplicity and to draw attention to particular gross discontinuities) and then summing over the adjacent pairs:



(a) Discontinuous

Adjacent pair	Ratio of diffs:* $\Delta y / \Delta x$	Squared ratio
a b	16/2	64.00
b c	2/11	0.03
c d	8/4	4.00
d e	2/9	0.05
e f	2/4	0.25

DISCONT (Sum) = 68.33



(b) Continuous

Adjacent pair	Ratio of diffs:* $\Delta y / \Delta x$	Squared ratio
a b	3/2	2.25
b c	11/11	1.00
c d	6/4	2.25
d e	7/9	0.61
e f	3/4	0.56

DISCONT (Sum) = 6.67

*i.e. $(y_i - y_j) / (x_i - x_j)$

Figure 5.2 *Discontinuous and continuous relations between two continua*

$$\text{DISCONT} = \sum_{\substack{j,k \\ \text{adjacent}}} \left(\frac{\Delta y}{\Delta x}\right)^2 = \sum \left(\frac{y_j - y_k}{x_j - x_k}\right)^2$$

This measure is calculated in the boxes alongside the two examples in Figure 5.2. The relatively continuous relation (B) has a value close to 5 (the minimum value of DISCONT), and the discontinuous relation (A) has a value of 68.33. (Note that in this latter case the value of DISCONT is most affected when small differences in x are accompanied by large differences in y, e.g. for (a, b) and (c, d). By contrast, where large changes in x give rise to small changes in y the contribution is very small, and this measure will be largely insensitive to them.)

5.2.2.1 Kappa as an index of continuity

The continuity transformation is used in MDS to obtain a solution where differences in the data correspond as smoothly as possible to differences in the solution differences. To do this we need the use of measure such as DISCONT, rather than stress. But in adapting DISCONT to measure the discontinuity between *multidimensional* spaces (rather than uni-dimensional continua) we run into a

problem. With a single line, the idea of a small change in value as we move up the continuum is easily defined: it is the difference between *adjacent* object locations. The notion of 'difference' generalises perfectly easily to 'distance' in the multidimensional case, but a little thought will convince you that there is no equivalent to 'adjacent' points in a two (and higher) dimensional space. But there is an approximation that will suffice: 'adjacency' can be replaced by 'closeness' or 'relative proximity' so long as we take care that only information relating to the immediate vicinity of each point is taken into account. In constructing an index of discontinuity in the multidimensional case, we shall therefore want to emphasise the distance involving closely proximate points and successively de-emphasise those at increasing distance. (This is obviously a further instance of *local* monotonicity described in the previous section.) In the context of MDS, the DISCONT measure is known as the 'kappa' index, symbolized by κ. The simplest measure on the analogy of stress, is referred to as 'raw kappa' and consists of two components, a discontinuity ratio and a weighting factor which restricts attention to the most proximate points:

$$\text{(raw) kappa} = \begin{matrix}\text{discontinuity} \\ \text{ratio}\end{matrix} \times \begin{matrix}\text{local proximity} \\ \text{weighting factor}\end{matrix}$$

$$\kappa = \left(\frac{\Delta x_{jk}}{\Delta y_{jk}}\right)^2 \times w_{jk}$$

Discontinuity ratio

In MDS applications we wish to ensure that small changes in the solution distances (d_{jk}) are associated with small changes in the data (δ_{jk}). Working with squared distances, as in DISCONT, the ratio becomes*:

$$\sum_{j=k}\sum (\delta_{jk}^2/d_{jk}^2)$$

Weighting factor

In the case of kappa, the weight factor is made the reciprocal of the corresponding squared solution distance:

$$w_{jk} = 1/d_{jk}^2$$

This form of weight has two useful properties: it ensures that local monotonicity is preserved (decreasing the contribution of any pair by the square of its distance, so proximate pairs contribute a good deal, and far distant ones scarcely at all), and the weights remain invariant under changes of scale.

Put together, these form the raw kappa index:

$$\text{Raw } \kappa = \sum_{j=k}\sum \left(\frac{\delta_{jk}^2}{d_{jk}^2}\right)\left(\frac{1}{d_{jk}^2}\right)$$

or, in simplified form:

$$\boxed{\text{Raw } \kappa = \sum_{j=k}\sum \left(\frac{\delta_{jk}^2}{d_{jk}^4}\right)}$$

*See Shepard and Carroll 1966, p. 581 et seq. In their treatment, data are referred to as (d_{jk}^2) and solution distances as (D_{jk}^2).

As in the case of raw stress, this index has the unfortunate property that an arbitrary enlargement of the solution configuration can make departure from continuity (raw kappa) as small as desired. And once again, the remedy is a normalising factor that will ensure that changes in the scale of the solution do not affect the index. Shepard and Carroll (1966, p. 583) show that the simplest effective normalising factor is:

$$\mathrm{NF} = 1 \Big/ \left(\sum\sum_{j \neq k} \frac{1}{d_{jk}^2} \right)^2$$

The normalised index of discontinuity (used in PARAMAP and non-linear PROFIT) then becomes:

$$\text{Normalised } \kappa = (\text{Raw } \kappa)/\mathrm{NF}$$

$$\boxed{\kappa = \sum\sum_{j \neq k} \frac{\delta_{jk}^2}{d_{jk}^4} \Big/ \left[\sum\sum_{j \neq k} \frac{1}{d_{jk}^2} \right]^2}$$

These, and related measures, are further discussed in Appendix A5.1 and in the PROFIT and PARAMAP documentation of the MDS(X) series.

By minimising kappa, continuity scaling both preserves local structure and allows solutions to be forced down into very small dimensionality, so long as the user is prepared to disregard or downgrade large distances. The Shepard diagram resulting from continuity scaling has a characteristic fan-like form which reflects these properties. As the (solution) distances increase, the corresponding data values increase, which reflects the fact that any discrepancy in the representation of small distances is heavily penalised (i.e. local structure is being preserved), whereas even very large discrepancies in representing the largest distances are virtually ignored. Typical examples of the diagram occur in Shepard and Carroll (1966, p. 575) and in Coxon and Jones (1978b, p. 266), reproduced as Figure 5.3.

Continuity scaling is a hybrid transformation. In that it assumes that the data are a direct estimate of the solution distances (except for a possible scaling factor), so it implicitly assumes that the data are at the ratio level of measurement, and is therefore an instance of classic metric scaling (see section 5.2.3.2). But it also preserves local monotonicity, and to that extent continuity scaling can be viewed as an even weaker form of monotonicity than that assumed by non-metric scaling. However, the continuity criterion ensures that the characteristic 'steppiness' and 'angularity' of the monotone function are smoothed out.

5.2.3 Regular transformations

By 'regular' transformations we mean those which are expressible in a simple mathematical form and are systematically increasing or decreasing. In effect, the term covers ratio and linear—often confusingly called 'metric'—and power rescaling functions. Regular rescaling transformations have the advantage over irregular monotonic transformations of being smooth and simple in form. Hence if the researcher's main interest focusses upon the relationship between the data and the underlying model, rather than on the solution itself (as, for example, in studying the relation between physical and perceived properties of colour or

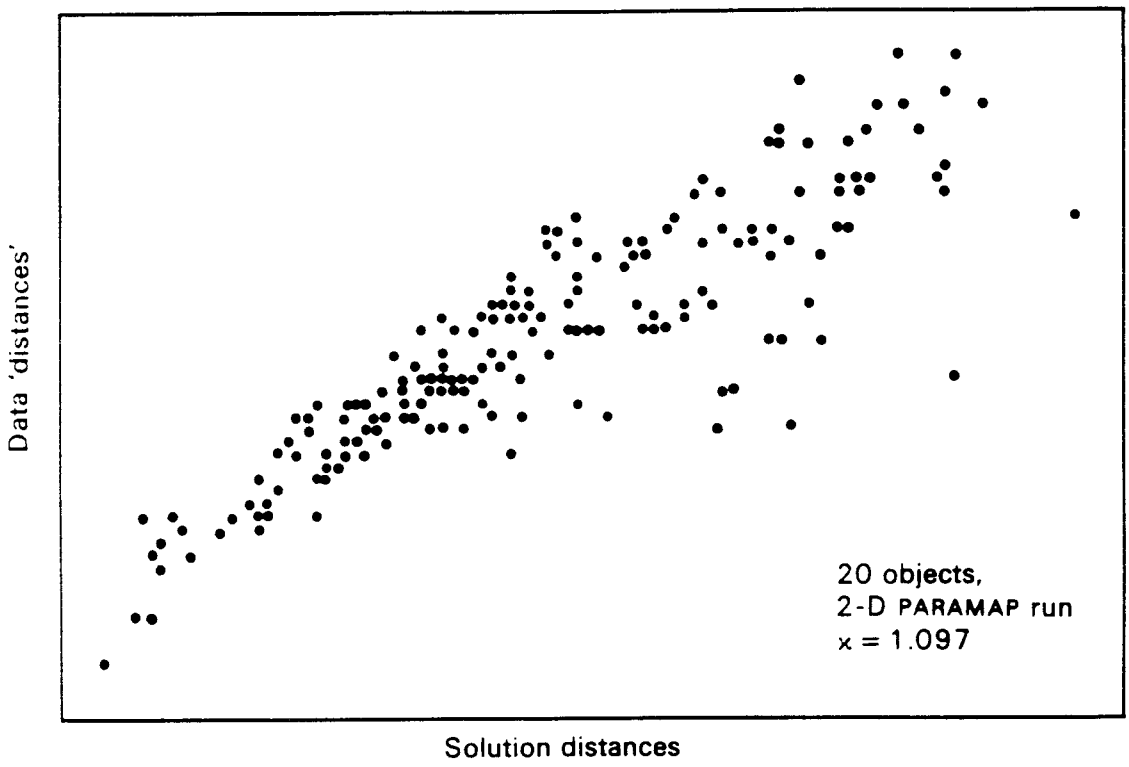


Figure 5.3 *Shepard diagram from continuity scaling*

between subjective and geographical distance), it is usually much simpler to interpret the results and predict values outside the current range of data if the transformation is a regular and simple mathematical function. In any event, a monotonic scaling often suggests a simpler, underlying relationship: Shepard functions are often linear or exponential over most of the range of the data. In such cases, having used the more indulgent monotonic assumption, it makes eminent sense to go on to use a more restrictive but simpler transformation and submit the data to metric scaling by a regular transformation.

5.2.3.1 *Ratio transformation*

The earliest forms of 'metric' multidimensional scaling, dating from the pioneering work of Richardson (1938), assumed simply that data dissimilarities were direct estimates of distances between the points concerned, so that the solution distances are viewed as a ratio transform of the distances of the solution, of the form

$$d_{jk} = b\delta_{jk},$$

where b is the 'proportionality coefficient' or 'scaling ratio', merely allowing for a difference in the actual size of the solution configuration, which is generally considered irrelevant in the MDS context. Given such a set of data, it is a relatively straightforward matter to estimate the dimensionality of the solution space and the co-ordinates of the objects by a method developed by Young and Householder (1941), known subsequently as Eckart-Young factoring (see Appendix A5.2).

5.2.3.2 *Linear transformation*

Linear transformations preserve information on the equality of intervals or differences, so that if the differences $(a - b)$ and $(c - d)$ are equal in the original data, they will also be equal when transformed linearly.

In many cases, methods of data collection or preliminary scaling yield quantities which clearly are not ratio-level genuine distances, but rather interval-level quantities sometimes referred to as distances. How are such interval-level data to be converted into ratio distances? The use of such distances as data assumes that, at least in the perfect case, the solution distances are a *linear* transformation of the data, that is,

$$d_{jk} = a + b\delta_{jk}.$$

In the usual case, this equation will only hold strictly for the fitted pseudo-distances, that is, $d^0_{jk} = a + b\delta_{jk}$. We have seen that the proportionality coefficient, b (the scale of the configuration) is arbitrary and merely chosen for convenience. However, estimation of the constant a (the intercept on the Shepard diagram linear regression function)* poses a more serious difficulty referred to as 'the additive constant problem'.

The additive constant problem

The problem can best be illustrated by an example based upon one originally given by Torgerson (1958, p. 403). Consider the matrix of data dissimilarities given in Table 5.2a. It happens that, as they stand, these dissimilarities cannot be represented in Euclidean space. The data do not even all satisfy the triangle inequality axiom of any distance measure (Appendix A2.1). For instance, the axiom requires that $d_{24} \leqslant d_{25} + d_{54}$, whereas in these data $d_{24}(= 6)$ is manifestly *greater* than $d_{25} + d_{54}(= 4)$.

If, however, each dissimilarity in Table 5.2a has a constant value of 2 added to it—that is, if the data are linearly transformed by the equation

$$\delta^{\text{new}} = 2.0 + (1.0)\delta^{\text{old}}$$

then the resulting data matrix is as given in Table 5.2b. It happens that there is a perfect two-dimensional representation of these data given in Figure 5.4. If, however, a constant greater than 2 is added, the data can still be perfectly represented, but only in a space of more than two dimensions.

The linear rescaling problem can be stated as follows: Given a data matrix which may not even be capable of representation in a Euclidean space, can a constant be found (i.e. how can the data be linearly transformed) so that the data can be represented as Euclidean distances (in as few dimensions as possible)?

There is no complete solution to the problem, though several have been proposed, some of considerable complexity (see Messick and Abelson 1956; Cooper 1971). An approach which has proved to be generally adequate is Carroll and Wish's (1973) 'triple equality' procedure (based upon Torgerson (1958, p. 276)) which converts data dissimilarities into distances by application of the 'triple equality difference' (TED) test to estimate the additive constant:

$$a = \max_{i,j,k} (\delta_{ik} - \delta_{ij} - \delta_{jk})$$

The 'triple equality difference' procedure is based upon a very simple idea. Let us

*In fact, MRSCAL estimates a slightly different transformation: $d_{jk} = b(\delta_{jk} + a)$, which results in a Shepard diagram where the function goes through the origin.

(a) *Data dissimilarities (relative or comparative distances)*

Object	1	2	3	4	5
1	—	3	4	3	1
2	3	—	3	6	2
3	4	3	—	3	1
4	3	6	3	—	2
5	1	2	1	2	—

$= \delta_{jk}$

(b) *Transformed data (actual distances)*

	1	2	3	4	5
1	—	5	6	5	3
2	5	—	5	8	4
3	6	5	—	5	3
4	5	8	5	—	4
5	3	4	3	4	—

$\delta'_{jk} = \delta_{jk} + 2$

(c) *Triple equality test on data of (a)*

Triple Points	*(Max)* (i, k) j	*Test* $(\delta_{ik} - \delta_{ij} - \delta_{jk})$	Result	
(1 2 3)	1, 3 2	4 − 3 − 3	− 2	
1 2 4	2, 4 1	6 − 3 − 3	0	
1 2 5	1, 2 5	3 − 1 − 2	0	
1 3 4	1, 3 4	4 − 3 − 3	− 2	
1 3 5	1, 3 5	4 − 1 − 1	+ 2	(max)
1 4 5	4, 5 1	2 − 3 − 1	− 2	
2 3 4	2, 4 3	6 − 3 − 3	0	
2 3 5	2, 5 3	2 − 3 − 1	− 2	
2 4 5	2, 4 5	6 − 2 − 2	+ 2	(max)
3 4 5	4, 5 3	2 − 3 − 1	− 2	

Additive constant = max $(\delta_{ik} - \delta_{ij} - \delta_{jk}) = 2$

Table 5.2 *Additive constant example*

suppose that the three points (i, j, k) form a straight line in the solution space, such as the line (1, 5, 3) in Figure 5.4a. Then $(d_{ij} + d_{jk})$ will necessarily be equal to d_{ik}, and hence the TED value, which may equivalently be written as $d_{ik} - (d_{ij} + d_{jk})$, will be zero. If j lies *off* the line then $(d_{ij} + d_{jk})$ will be larger than d_{ik} and hence the value of TED will be *negative*. In short, the TED test applied to a set of actual distances will produce a value of 0 for points lying on a line, and a negative value in other cases. Note that in this case the test could here never have a positive value and its maximum value would be zero. The situation is the same when dealing with data or 'relative distances' (where $\delta_{mn} = d_{mn} + a$) except that the TED test will give rise to the value $(0 + a)$ in the case of collinear points and to a smaller value (negative $+ a$) in other cases. Hence the *maximum* value of TED will give the quantity which has to be added to each dissimilarity value to convert it to a genuine distance, i.e. the 'additive constant'. This number may incidentally be

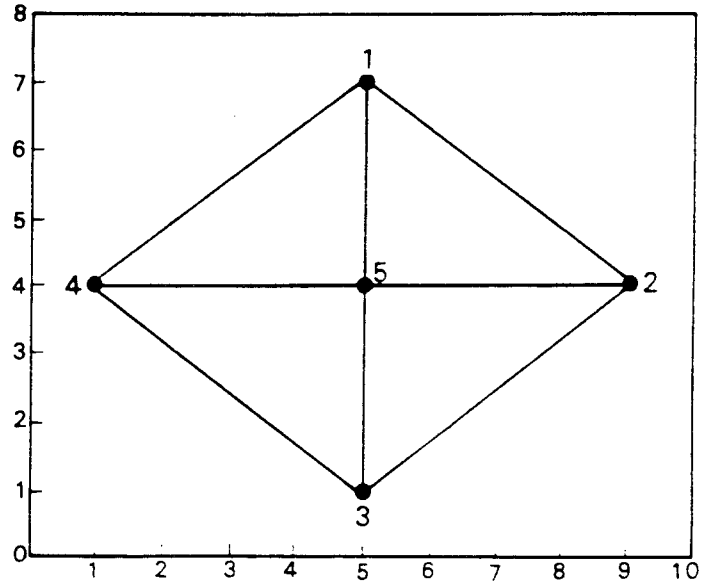


Figure 5.4 *2-dimensional representations of data in Table 5.2b*

negative. As an example, consider the data in Table 5.2. In Table 5.2c an additive constant of 2 is necessary to turn the data into real distances. This value is correctly given by the triples of points (1, 3, 5) and (2, 4, 5), and in both cases the three points lie as a straight line, as can be seen in Figure 5.4a. Even for fallible data, and so long as there are enough points to ensure that at least some triples come close to forming a straight line, this simple method provides an adequate and straightforward way of estimating the additive constant, and is the method used in the INDSCAL program.

5.2.3.3 Power (and log-interval) transformations

Power transformations have the general form: $x' = kx^{\beta}$ and preserve information not only on the equality of intervals—as in the interval scale—but also on the equality of *relative* intervals, i.e. on the ratio of data values. For instance, taking four ratio level data, $a = 3$, $b = 6$, $c = 10$ and $d = 20$, then the ratios a/b and c/d both equal $\frac{1}{2}$. When the values are transformed by the power function $x' = 3x^2$, the ratios a/b and c/d are still equal, but now equal $\frac{1}{4}$. In the equation the value of k is an arbitrary factor which cancels out in the formation of ratios; it is the exponent, β, which carries the significant information. Power functions are probably most familiar in the form of compound interest rates in economics and in the 'psychophysical law' in psychology.

A power relationship can always be re-expressed in logarithmic form.* In logarithmic form, power transformations preserve the log *differences* (or intervals) corresponding to the original ratios so that, in the above example, $\log a - \log b = \log c - \log d$ whether with the original values or under the transform $x' = 3x^2$, as can easily be checked. For this reason, the power transformation is sometimes called the logarithmic interval scale, the term adopted in this context by Stevens (1959, pp. 29–30) and Roskam (1972, pp. 495–506). The power transformation is implemented in logarithmic-interval form in the MRSCAL program.

*$10^2 = 100$, and $\log_{10} 100 = 2$, and, in general, if $a^b = c$ then $\log_a c = b$.

The power transformation is a smooth, regular, but non-linear function, illustrated in Figure 5.1(i), whose main parameter of interest is the exponent value which determines how rapidly the slope accelerates. If the power function is drawn in log co-ordinates it then appears as a straight line, with slope equal to the value of the exponent, β. Put in log-interval form, the power transformation† in the case of perfect data would be

$$d_{jk} = a + b(\log(\delta_{jk})).$$

where a represents an additive constant (which may have psychological meaning as the threshold value—see above—but is not usually given substantive interpretation) and b represents the exponent value.

The power transformation has received considerable attention in scaling because of its centrality in early psychological studies of the relationship between physical variables and their subjective counterparts, and also because some data and judgmental processes are known to be best represented by such a transformation.

The 'power law' and its scaling consequences

The work of Fechner and Weber from the 1850s on suggested that human subjects noticed a change in the intensity of a physical variable (such as sound pressure) when the change represented a fixed *proportion* of the previous intensity, i.e. that a *relative* increase in a physical property was perceived as a unit *fixed* increase in psychological intensity. Put slightly differently, the subjective intensity increases as a power function of the physical intensity. Later research has shown that for a wide variety of physical properties, the relationship is well approximated by the so-called psychophysical law (Stevens 1974, p. 361)

$$\psi = k\phi^{\beta},$$

where ψ is the perceived magnitude or intensity, ϕ is the physical magnitude, β is the power exponent and k is an arbitrary scaling factor. (In some cases the psychological magnitude only begins to be experienced at a particular threshold and in this case the form of the 'law' needs slight alteration by including an additive constant to represent the threshold effect. Substantive interest focuses the typical value of the power exponent (β) for various modalities.‡

Later experimentation has suggested that a very similar power relationship also exists for the intensity of opinions and attitudes and for the relationship between *direct* estimation (rating) of attitudinal areas and their *indirect* measurement, derived by such methods as Thurstone's law of comparative judgment (Stevens 1966 provides a wide range of examples).

If the 'power law' holds for 'softer', non-experimental and more complex phenomena, as Stevens and others argue it does, then some important consequences follow for scaling studies.

First, 'objective' external properties may well be non-linearly related to scaling solutions based upon subjective or perceptual data. At the very least, it would be

†As in the linear case, the MRSCAL program actually estimates: $d_{jk} = b(\log(\delta_{jk}) + a)$.

‡Each modality tends to have characteristic exponent values, ranging from $\frac{1}{3}$ for brightness, $\frac{2}{3}$ for loudness to $3\frac{1}{2}$ for the subjective intensity of electrical current. See Stevens (1974, pp. 362 et seq.).

prudent to allow for this eventuality when engaged upon property-fitting using PROFIT, allowing a 'continuity'-based relationship which will tend to keep increments, and hence ratios, fairly constant, or allowing a monotonic—and hence a power—relationship between the property values and the configuration distances using PREFMAP. In either case it would be foolish only to choose the linear option, which would badly distort a genuine power relationship.

Secondly, the assumption of linearity between the data and the solution is likely to be highly suspect if the data collection method was 'direct' rather than 'derived' (see 2.2 and 2.3). Thus, if the linear transformation is used, it ought to be supplemented by a monotonic fit and/or a 'power' fit, and the Shepard diagrams should be compared.

Thirdly, another way of expressing the power law is that error or variability increases with the magnitude of the data. It is an important consideration in studies of consensus in human judgments (Stevens 1966) and in the development of more recent MDS models, e.g. Ramsay's 'multiscale models', which make explicit assumptions about the likely characteristics of error in the subject's data (see Ramsay 1977, pp. 243–6, especially the discussion beginning with the second paragraph of p. 245, and our section 8.2.1). Perhaps more to the point, if error increases with magnitude it is sensible to pay little attention to dissimilar points in obtaining an MDS solution. This provides a further reason for choosing the local monotonicity or continuity options.

Finally, for some types of data—and especially for confusion data, where the similarity between two objects is taken to be a function of the frequency with which they are confused—there are good theoretical and empirical reasons for expecting an exponential decay (negative power) relationship between the data and the solution distances. Indeed, this same characteristic J-shaped curve has been noted for a goodly number of non-metric scaling studies of co-occurrence frequency data, including Figure 3.14b, and it has been shown that the adoption of a power transformation for the MDS analysis in these circumstances often restores significant local structure which is lost in an ordinal scaling (Arabie and Soli 1977).

5.3 Models

The basic MDS model represents data values as distances. These distances may be thought of as being produced by the combination of latent parameters, i.e. the co-ordinates of the space, which might reasonably be interpreted as scale values along each dimension. It is the particular form of the composition of these co-ordinates to form distances which makes their interpretation as scale values problematical, for we are asserting in the distance model that the scale values for the stimuli are compounded into distances by taking the difference on each scale, squaring it, then summing over each dimension and finally deflating its value by taking the square root. It is possible, however, to regard the data as being linked to a set of scale values or co-ordinates by composition rules other than those of the Euclidean distance formula—and indeed, by variants of the distance formula.

Three major types of composition rule are usefully distinguished:

(i) *Simple composition.* Each category of each way of a two-way (or higher) table of data has a scale value, and the composition rule specifies that the entry is

the simple sum of the component categories (the additive model). Other commonly occurring examples are the difference (a subtractive) and the multiplicative (product) compositions.

(ii) *Scalar product (or factor) composition.* The objects are located as points and or as vectors in a space, and it is the angular separation (scalar product) of the vectors which corresponds to the data dissimilarities.

(iii) *Distance composition.* The objects are located as points in a space and the distance between the points represents the data dissimilarities.

Let us take each type of model in turn.

5.3.1 Simple composition

Quite frequently in empirical research the value of a dependent variable is considered to have been produced by the conjoint effect of two or more independent variables or factors, and the researcher is interested in estimating what the numerical effects are (the scale values) and how they combine. Examples abound: factorially-designed experiments in agriculture investigate how, and to what extent, different combinations of soil and fertiliser affect crop yield; social psychologists interested in impression-formation construct combinations of traits and ask subjects to rate the attractiveness of the resulting combinations; economists construct portfolios of investments or commodity bundles and ask respondents to give their preference orderings; demographers calculate the mean fertility of couples from different regions and occupational groups. In each case the basic notion is the same: the data are assumed to represent the simultaneous, conjoint effect of the defining factors, and the purpose of the analysis is to assign a scale value (estimate a numerical weight) to each constituent category of each independent variable, which, when combined according to the composition rule of the model, will best fit the values of the dependent variable.

Most researchers will have encountered this type of analysis in the context of two (and higher) way analysis of variance and the log-linear analysis of contingency tables. In both cases, the underlying model is an additive one: the values in the table of the dependent variable are assumed to be the *sum* of the effects of the relevant categories which define the entry. Given the following 2-way table of data, the scale values $A = (8, 2, 5, 4)$ and $B = (3, 1, 6)$ combine additively to produce the entries in the table, i.e. $x_{ij} = a_i + b_j$. In most actual applications, an additive model will not fit the data perfectly and further interaction terms may need

		B		
		b_1	b_2	b_3
A	a_1	11	9	14
	a_2	5	3	8
	a_3	8	6	11
	a_4	7	5	10

to be included to represent the unique, joint effect of the categories. However, it might be that a transformation—a rescaling—of the data *will* fit an additive model.

In a classic paper, Box and Cox (1964) discuss a number of polynomial transformations of the data, designed to render effects as additive as possible, and in his famous paper Kruskal (1965) developed a procedure based upon monotone regression designed to find an *ordinal* rescaling of such data which makes them maximally conform to an additive model. The affinity with the basic non-metric MDS model will be obvious, and Kruskal's procedure forms the basis of the additive sub-model of the UNICON program discussed later in the book.

So far we have implicitly assumed that the data form a 2-way table, for purposes of simplicity. Most MDS implementations of simple composition scaling allow up to five such ways, or factors (which may or may not be 'modes', i.e. not distinct sets of objects), although there are few empirical instances of anything more than 3-way table analysis.

Three basic operations form the basis of simple composition models:

(i) *additive model*: $x_{ij} = a_i + b_j$
(ii) *difference (subtractive) model*: $x_{ij} = a_i - b_j$
(iii) *multiplicative (product) model*: $x_{ij} = a_i \times b_j$

The *additive model* is undoubtedly the best-studied and most used. It turns out to be possible to formulate the necessary and sufficient qualitative conditions that a table must satisfy if it is to be capable of an additive representation. This constitutes a major triumph of axiomatic representationist measurement theory (Krantz et al. 1971, p. 423 et seq.).

The subtractive model (or difference model) is appropriate where, for instance, subjects have been instructed to judge the difference between pairs of objects (for example 'imagine 2 different people, each described by one of the adjectives of each pair, and then judge the difference in likeableness between the 2 persons', or where effects are expected systematically to counteract each other).

The product model is appropriate where it is thought that categories have a multiplier effect upon each other (or, equivalently, when the logarithm of the effects are additive).

The UNICON program allows the user to define a number of more complex models involving the three simple operations, such as:

$$x_{ijk} = a_i \times b_j + c_k + d_l$$

and

$$x_{ijk} = a_i + b_j - c_k.$$

(See program documentation for details.)

5.3.2 Scalar product models

In the MDS(X) series, all the scalar products (or vector or factor) models assume that the data consist of (or can be reduced to)* a rectangular *two-mode* matrix consisting of a set of (preference) ratings or rankings of a set of p stimuli made by a

*In the MDPREF vector model, input may be a set of pair comparison dominance matrices.

set of N subjects. (In MDPREF this matrix is termed the 'first score matrix'.) For convenience the entries in this matrix are usually denoted s_{ij} to mean the similarity between subject i and object j, or more usually the preference score given by subject i to object j.

The vector solution consists of a configuration of p stimulus points in a user-chosen number of dimensions, and each of the N subjects' set of preference ranks or ratings is represented as a vector, located so that the projections of the stimuli on the vector are in maximum agreement (correlate as highly as possible) with that subject's preferences. The external form of this analysis, i.e. where the stimulus configuration is obtained separately and remains fixed whilst the subject vectors are estimated, was discussed in section 4.4.1.

The purpose of these models is to represent both the stimuli and the subjects in a common 'joint space'. Each subject's preferences are represented as a vector—a projection down, or collapsing of, the stimulus space onto a single dimension—just like the properties embedded in a stimulus space. Interest will chiefly focus therefore on two things:

(i) how well the subject's preferences can be accommodated by the model, and hence represented in the stimulus space (this can be assessed by the correlation of the projections with the original data) and

(ii) how the vectors relate to each other, since the main purpose may be to investigate individual differences in a set of rankings/ratings.

Differences between rankings are signalled in the vector model principally by angular separation. On the one hand, as we saw earlier, the direction in which a vector points is highly significant, for it indicates the manner in which the subject mixes or trades off the characteristics of the stimuli in producing her preferences, and this is measured by the cosine of the angle which the vector makes with the dimensions of the space. By the same token, if we are interested in how one subject vector relates to another, we inspect the angular separation between them—the linear correlation, or cosine of the angle between the two vectors. In inspecting a vector model solution, the first point of interest is how the subject vectors are dispersed around the unit-circle (or sphere).*

If the vector ends are located in a small sector, this indicates high consensus or agreement in subjects' preferences, whereas the more unevenly they are distributed round the circle, the greater the dissensus. The researcher will presumably become interested in whether distinguishably different 'points of view' exist, suggested by small sectors with a high density of vector ends and empty sectors between sectors. If there are different categories of subjects we may also want to know whether the average direction differs significantly between the categories, and statistical tests and procedures for analysing directional data have been developed and are available. (They are discussed in Mardia 1972, and in the MDS context in Coxon and Jones 1979, pp. 128–36 as well as in the MDS(X) documentation for the MDPREF program.)

*By convention, subject vectors are normalised to have the same (unit) length in MDPREF. Though this is not a necessary restriction of the model, it makes for greater simplicity if vectors are of standard length. In two dimensions, vector ends will therefore lie around a unit circle, in three (and higher) dimensions, they lie around a (hyper) sphere.

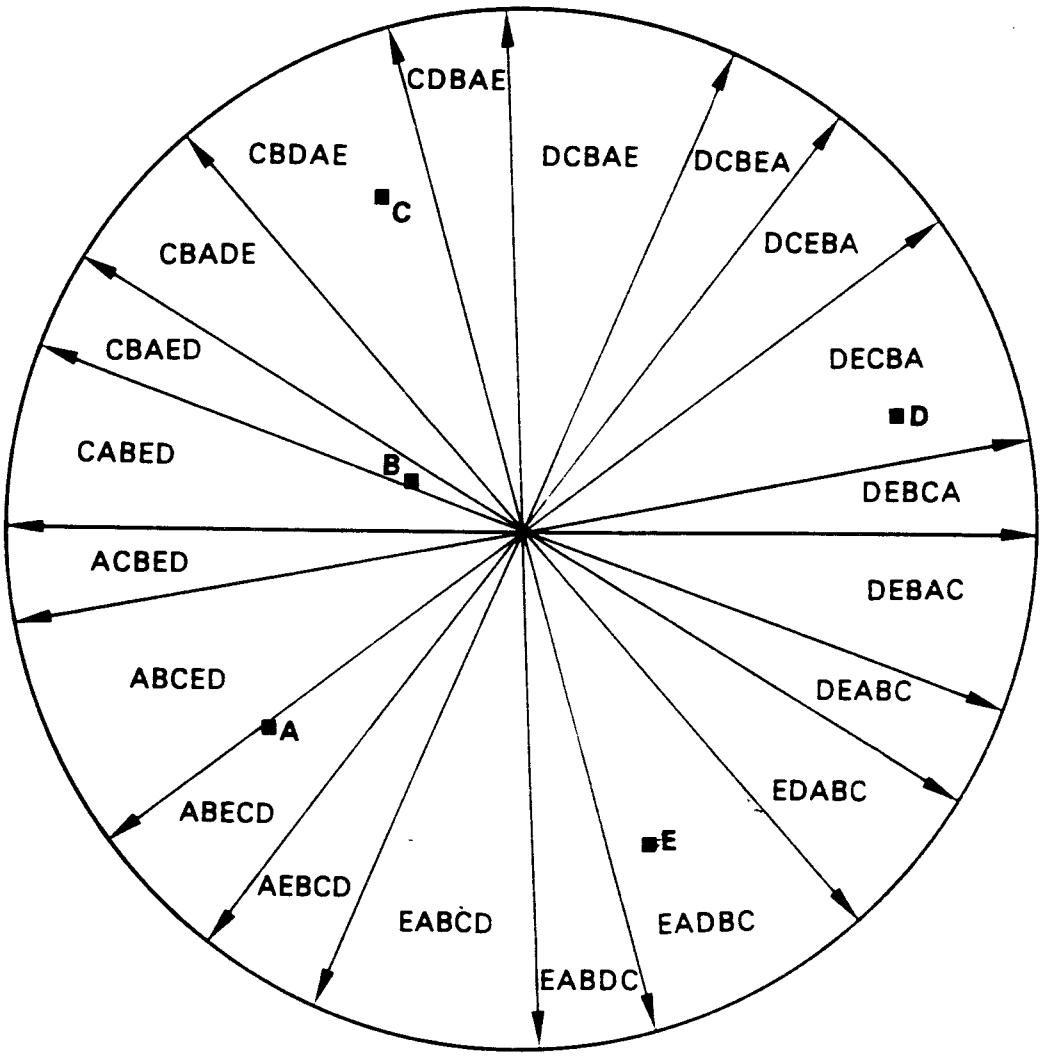
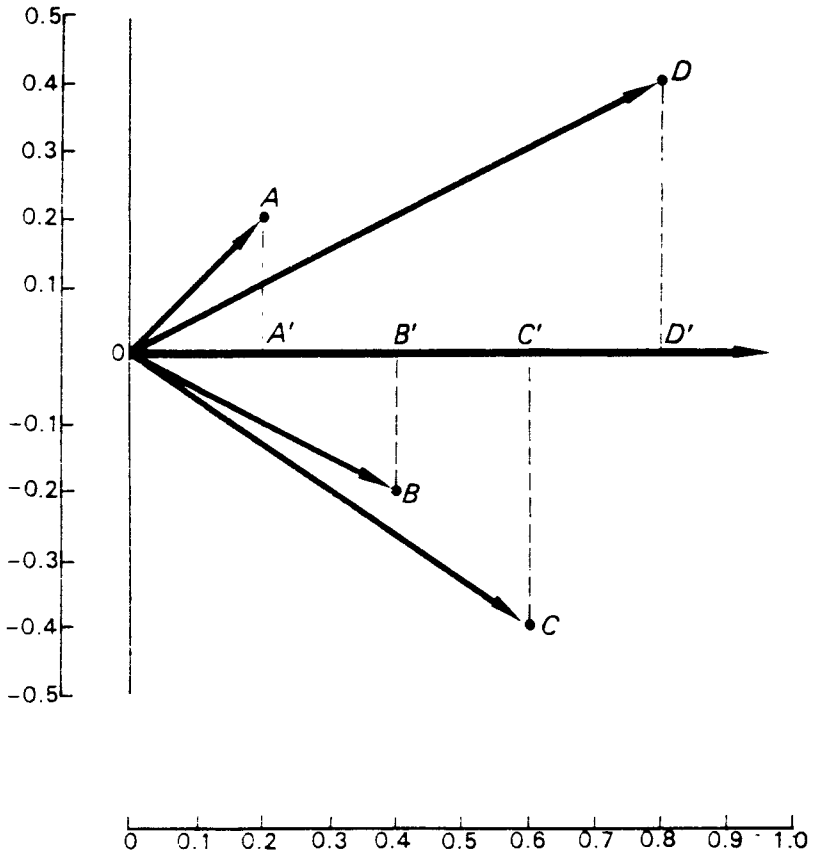


Figure 5.5 *Twenty rankings compatible with 2-D stimulus configurations of 5 points (vector models)*

Although a total of $p!$ (i.e. $p(p-1)(p-2)\ldots\times 1$) rankings of p objects is possible, only a limited number of these can be accommodated within a stimulus configuration. We therefore need to enquire both how many rankings can be accommodated in a configuration of p points in r dimensions and how they are represented therein. As an example, take the 5-point stimulus configuration given in Figure 5.5. There are $5! = 120$ possible rank orderings of 5 stimuli, but only 20 of these can be represented perfectly in a given vector configuration of 5 points in two dimensions, and one half of these will simply be mirror-images of each other formed by reversing the direction of the vector. The 20 rankings compatible with this configuration are given in the figure. Notice that there is an orderly interlocking between the rankings, akin to that shown by Coombs (1964, p. 87 et seq.) in the context of discussing the unidimensional unfolding (distance) model for preferences. As one moves around the circle, only adjacent stimuli are interchanged in the rankings (beginning in the north-easterly position and moving clockwise: *DCBEA*, *DCEBA*, *DECBA*, *DEBCA*, and so forth).

Although the scalar products model has been described as a point (stimulus) and vector (subject) representation, formally the model is expressed entirely in terms of vectors—a set of vectors drawn from the origin of the space to the location of each



Subject vector **y**	Stimulus point **x**	Scalar product = **yx'**	Projection
(1,0)	*A* (0.2, 0.2)	0.2	*OA'* = 0.2
	B (0.4,-0.2)	0.4	*OB'* = 0.4
	C (0.6,-0.4)	0.6	*OC'* = 0.6
	D (0.8, 0.4)	0.8	*OD'* = 0.8

Figure 5.6 *Projections and scalar products*

stimulus, and a set of unit-length *subject* vectors. The key to understanding the formula for the model is knowing that the scalar product of the stimulus vector with the (unit-length) subject vector is the same as the vertical projection of that stimulus point on the subject vector. This property is illustrated in Figure 5.6, where the subject vector is drawn along the first dimension to simplify the arithmetic.

Let $\mathbf{x}_j$ represent the vector from the origin to the location of stimulus j in r-dimensional space, and $\mathbf{y}_i$ represent the (unit-length) vector for subject i, then the preference value which stimulus j has for subject i is estimated as the scalar product of the vector concerned:

$$\hat{s}_{ij} = \mathbf{y}_i \cdot \mathbf{x}'_j = \sum_{a=1}^{r} y_{ia} x_{ja}$$

or in matrix form:

$$\hat{\mathbf{S}} = \mathbf{Y}\mathbf{X}'$$

The matrix of preference scores estimated by the model is termed the 'second-score matrix', and the purpose of the vector model is to obtain a stimulus configuration **X** and subject vectors **Y**, so that the discrepancy between the original 'first-score' data (s_{ij}) and the estimated 'second-score' values ($\hat{s}_{ij}$) is as small as possible. (In the case of a non-metric version, the monotonically transformed data will be compared to the estimated values.) Carroll (1972, p. 124 et seq.) and the MDS(X) documentation describe the stress-like index of agreement, C_1 used to measure the goodness of fit. The method of solution involves factoring two product matrices formed from the first score matrix.*

The main properties of the vector model (cf. Roskam 1968, p. 28) may be summarised as follows:

(i) *Increasing utility*. A subject's preference (or similarity rating) increases continuously in the direction of the vector: the further out an object projects on it, the more it is preferred.

(ii) *Mediocrity*. An object may always occupy a position between the extremes of all the subject's preferences, i.e. never be either most or least preferred (see object B in Figure 5.5, for example).

(iii) *Reversability*. If a given ordering occurs, the opposite ordering may also occur. Indeed, the orderings compatible with a given stimulus configuration divide into two opposite halves, producing the characteristic 'spokes of a wheel' isotonic regions (sector of the space where the same rank ordering of stimuli is implied) seen in Figure 5.5.

The vector model differs considerably in these respects from the distance (unfolding) model of preference discussed in the next section. The differences and the related issues of interpretation of configurations produced by programs implementing the models are discussed in Chapter 6.

5.3.3 Distance models

The central idea of distance models is that the proximity of points in a space is used to represent their empirical similarity, or equivalently that distance represents their dissimilarity. In the vast majority of MDS models, the distance function involved is the familiar Euclidean form, but Euclidean distance is only one special case of a whole family of distance functions, each with its own characteristics and properties (see Appendix A2.1.1.2). Proceeding from the familiar to the less familiar, we shall discuss the basic distance model first, then move on to look in greater detail at the properties of Euclidean and other types of distance.

Given a set of distances it is always possible to reconstruct the configuration of points which generated them. (This procedure is described in Appendix A5.2.2 and forms the basis of classic metric scaling discussed above.) However, such a recovered configuration is not unique, in that several aspects of it are arbitrary and

*The first score matrix **S** is approximated in the user-chosen dimensionality a, by a least squares approximation $\mathbf{S} = \mathbf{YX}'$ (of rank a) using the Eckart-Young factorising procedure. The eigenvectors of the minor product matrix **S'S** provide estimates of the stimulus configuration **Y**, and the eigenvectors of the major product matrix **SS'** provide the estimates of the subject vectors **X**, when the rows are normalised to unity. The eigenvalues of both product matrices are the same and indicate the concentration of variation in the principal axes (see Appendix A5.2.2).

may be changed at will. (These have been mentioned before (4.1), and are further discussed in Appendix A7.1.) In particular, the actual size or scale of the configuration and the origin of the space are arbitrary. Moreover, the orientation of the axes may be changed and reflected at will. Strictly speaking, it is only the relative distance between points which is significant in interpreting a distance model solution—the origin and axes simply provide a convenient framework to locate the points.

5.3.3.1 Point-point ('two-mode 'unfolding') distance models

When the data consist of a rectangular two-mode matrix, of rankings or ratings, then the distance model can be used to represent both the stimuli *and* the subjects as points. The solution consists of a configuration of p stimulus points and N subject points where each subject is represented as being at a 'maximal' or 'ideal' point, located in such a way that the distances from this point to the stimulus points are in maximum agreement with the subject's preference ratings or rankings.

In external models such as PREFMAP phase III, the stimulus configuration is obtained separately and remains fixed whilst the 'subject' or property points are estimated (see 4.4.2), whereas in internal models, such as MINI-RSA, both sets are estimated simultaneously. As in the case of the vector model, both metric and non-metric versions exist—in the former a linear correlation between the preference data and the subject-stimulus distances is maximised while in the latter a variant of stress involving only the rank order of the data is minimised.

The position of the 'ideal point' is interpreted as the one point in the space where the subject's preferences are at a maximum, and her preference decreases in every direction. This is often termed a 'single peaked preference function', since it assumes that there is only *one* point of maximum preference.

The non-metric version of the distance model is best known under the title of 'unfolding analysis', developed by Coombs (1964, chs. 5–7). The two-dimensional case is illustrated in Figure 5.7 with reference to the same 5-stimulus configuration used in the vector model case (Figure 5.5).

A midline is drawn between each pair of points, dividing the space up into 46 isotonic regions. Every ideal point within one of these regions possesses the same rank order of distances to the five stimuli. This is illustrated in Figure 5.7; thus in region I the corresponding I-scale is ABECD, and in crossing over the midline CE to region II, the I-scale becomes ABCED. Similarly the move from region III to IV represents the transition from DBCEA to DBECA. Notice that some regions are entirely encompassed by midlines (closed isotonic regions), whilst others at the periphery are not (open isotonic regions). Herein is an important distinction between the vector and distance models: the vector model excludes closed regions (see the corresponding Figure 5.5) and can accommodate fewer I-scales than the distance model. The maximum number of I-scales compatible with the two models is illustrated below in Table 5.3 (see Coombs 1964, Tables 7.1 and 12.9).

Normally the points corresponding to the most popular or consensual rankings will lie at the centre of the space, and the least popular ones at the periphery. Research has shown, as Coombs originally suggested, that ideal points within the 'open' isotonic regions are located with less accuracy than those in the closed ones. Moreover, the fewer the midlines constraining a region, the more likely it is that the

No. of points	*No. of dimensions*								
	2		3		4		5		*Total possible (p!)*
1	1	*1*	1	*1*	1	*1*	1	*1*	1
2	2	*2*	2	*2*	2	*2*	2	*2*	2
3	6	*6*	6	*6*	6	*6*	6	*6*	6
4	18	*12*	24	*24*	24	*24*	24	*24*	24
5	46	*20*	96	*72*	120	*120*	120	*120*	120
6	101	*30*	326	*172*	600	*480*	720	*720*	720
7	197	*42*	932	*352*	$2.56_{10}3$	*1512*	$4.32_{10}3$	$3.60_{10}3$	$5.04_{10}3$
8	351	*56*	$2.31_{10}3$	*646*	$9.08_{10}3$	$3.98_{10}3$	$2.22_{10}4$	$1.42_{10}4$	$4.03_{10}4$
9	583	*72*	$5.12_{10}3$	$1.09_{10}3$	$2.76_{10}4$	$9.14_{10}3$	$9.49_{10}4$	$4.60_{10}4$	$3.63_{10}5$
10	916	*90*	$1.04_{10}4$	$1.74_{10}3$	$7.36_{10}4$	$1.90_{10}4$	$3.43_{10}5$	$1.28_{10}5$	$3.63_{10}6$
MODEL: DISTANCE/*VECTOR*	D	*V*	D	*V*	D	*V*	D	*V*	

Table 5.3 *Total possible number of I-scales, and totals compatible with the distance and vector models for* p *points in* r *dimensions)*

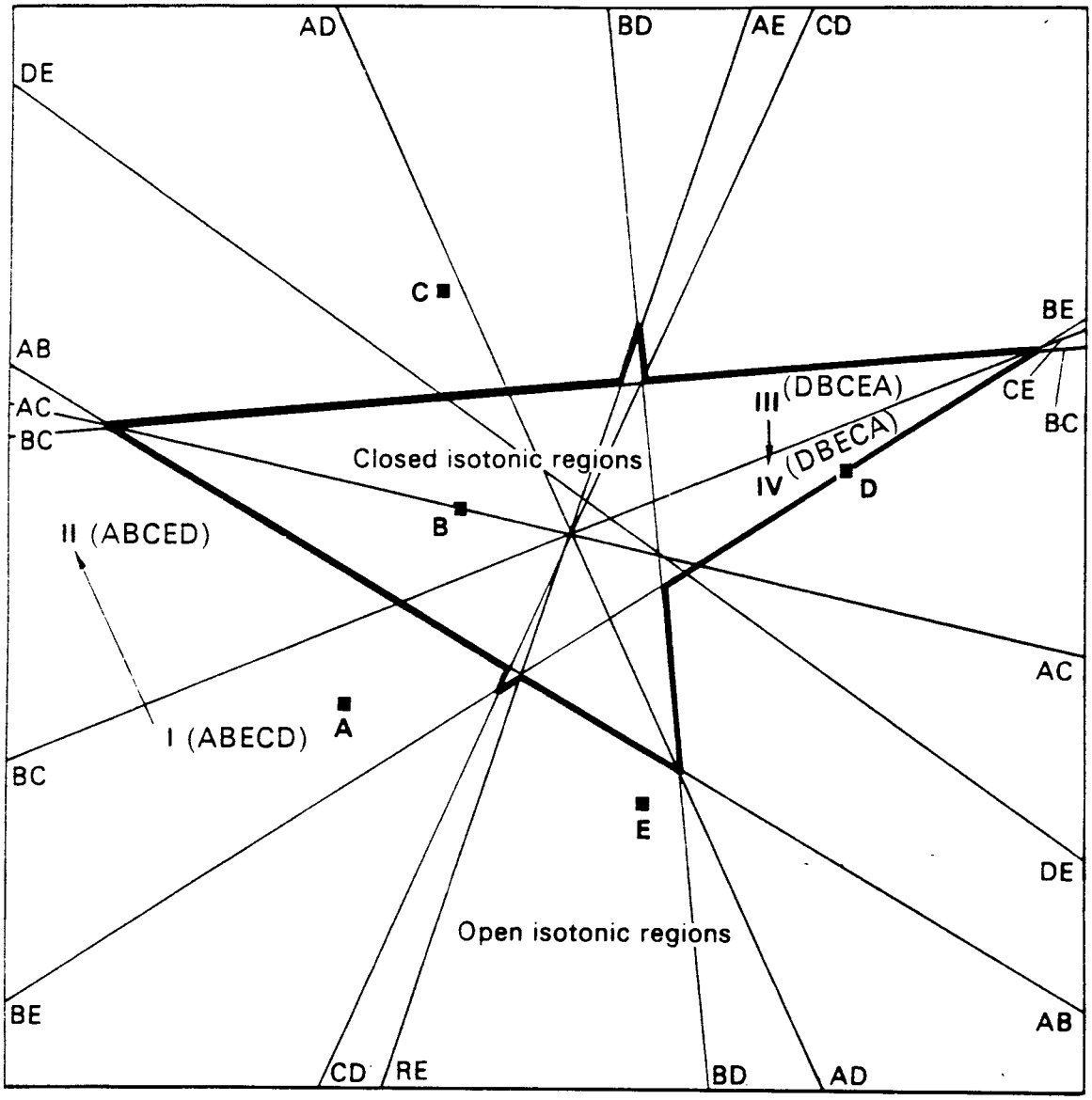
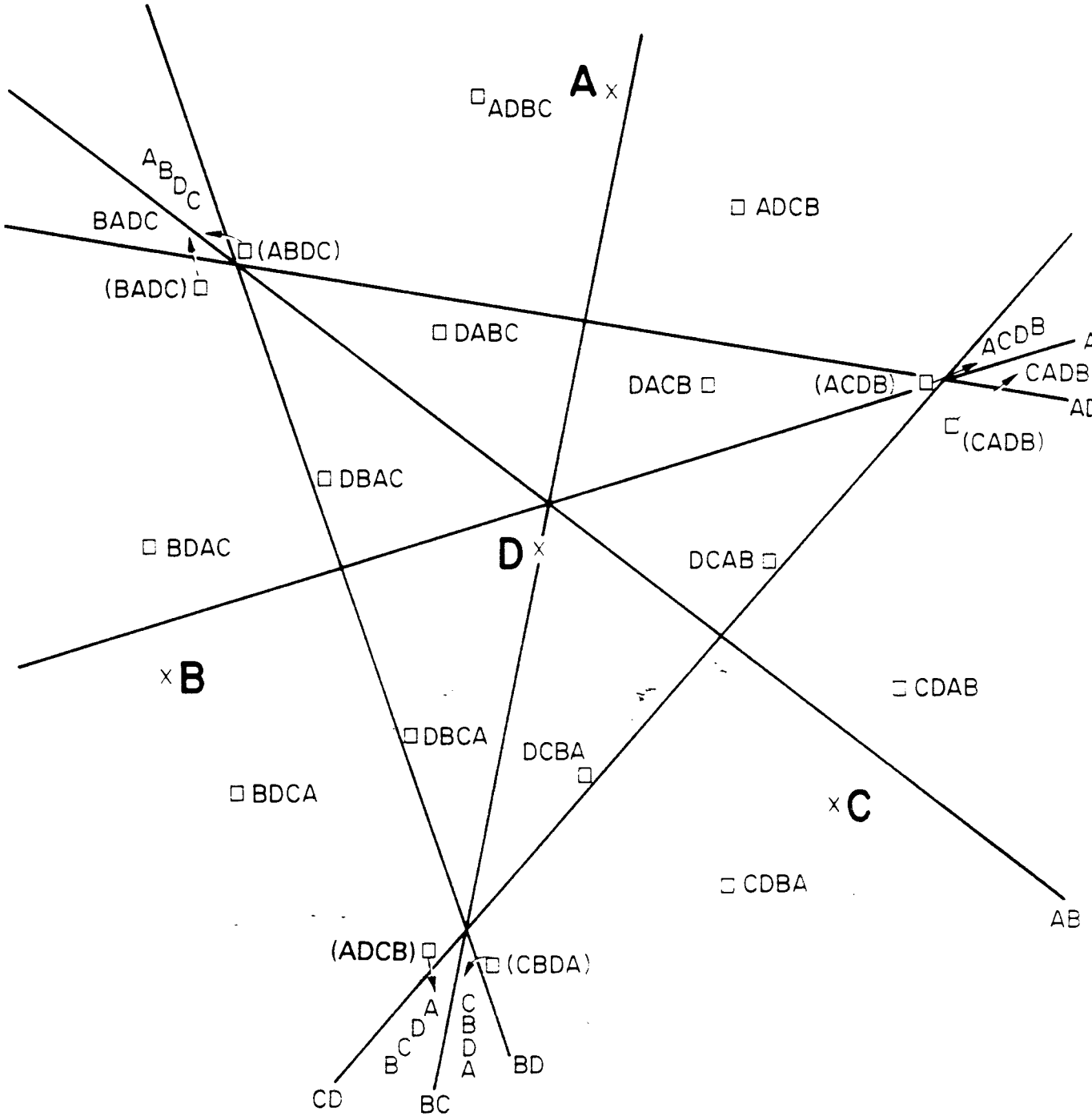


Figure 5.7 *Forty-six rankings compatible with 2-D stimulus configuration of 5 points (vector model)*

subject point be mislocated in a scaling solution. This is well illustrated in Figure 5.8, representing the scaling of the 18 I-scales compatible with a 2-dimensional, 4-point configuration (Coombs, 1964, Figure 7.4, p. 146). Each small square represents the position of a subject as located by the relevant non-metric program (MINIRSA). Note that in the case of the closed regions, the squares are all located within the correct region, although they are deflected to the outer edge. In the case of the open regions, those defined by three lines are correctly located near the centre of the region but those defined by only two lines are, without exception, displaced slightly outside their correct location.

The multidimensional unfolding model is hence clearly more 'tolerant' than the vector model, in the sense that it can accommodate more I-scales (see Table 5.3). So long as the number of stimulus points is large compared to the number of dimensions, the size of the isotonic regions is small, especially towards the centre of the configuration, and they become increasingly well-represented by a point. For this reason, stimuli points in the central part of a configuration are normally the most stable, whilst those at the periphery can usually be moved around fairly freely without affecting the goodness of fit. The variation in judgments about particular



$Stress_2$ = 0.0003 after 45 iterations

N.B. Arrows indicate discrepancy between 'true' locations and scaled ('recovered') location

Figure 5.8 *Actual and scaled location of isotonic regions*

stimuli is also an important factor in assessing the stability of a configuration in an internal scaling model. Highly popular stimuli will tend to be projected into the centre of the subject points (so that they can feature close to most subject's ideal points) and highly unpopular stimuli will be located at the outside of a configuration. Indeed, if a stimulus is sufficiently unpopular it can be located virtually anywhere on the periphery, so long as it is at a maximum distance from the ideal points. An example of this occurs in the analysis of the Delbeke data reported in section 6.2.2 (see Coxon 1974) where virtually all subjects rejected the

stimulus 'no children' in a study of preferences for families of different sizes and sex composition. When scaled, this stimulus was located at greatly varying points, but always at an extreme distance from the centre.

In summary, the properties of the point-point (distance) model of preference which contrast with the vector model are as follows:

(i) *Single peakedness.* It is assumed that each subject has one single point of maximum preference and that preference decreases (symmetrically) from this point.

(ii) *Excellence.* If the distance model holds, then each stimulus must be preferred most by at least one subject.

(iii) There is nothing corresponding to the reversability property of the vector model in the multidimensional unfolding model: some mirror-image pairs of I-scales will exist, but not others. More importantly, the distance model is characterised by the presence of closed isotonic regions, which cannot occur in the vector model.

5.3.3.2 Euclidean and non-Euclidean distance

So far, 'distance' and 'Euclidean distance' have been used interchangeably. In fact, a whole family of distance measures can be defined for a given configuration of points. Our interest shifts away from the correct location of points to how we measure the distance between them.

Three types of distance have been found useful in MDS and are represented in various MDS(X) programs: city block, Euclidean and dominance metrics. These are all special instances of the Minkowski *r*-metric family of distance measures which have the form:

General (Minkowski) Distance

$$d_{jk}^{(r)} = \sqrt[r]{\sum_a |x_{ja} - y_{ka}|^r}$$

where x_{ja} is the co-ordinate of the k th point and y_{ka} is the co-ordinate of the j th point on the a th dimension and r is the Minkowski r-metric power.

Each value of r (between 1 and infinity) defines a distinct metric distance. Each can be thought of as a simple composition model—a 'powered additive difference' model which asserts (Beals et al. 1968 pp. 133–5) that:

(i) absolute *differences* on each dimension, a
(ii) which are raised to the same *power* r
(iii) combine *additively* over the dimensions to produce
(iv) the overall distance between a pair of points, j and k.

In the case of Euclidean distance, the power is 2, so differences are squared, and the final distance measure deflates the value by taking the square root.*

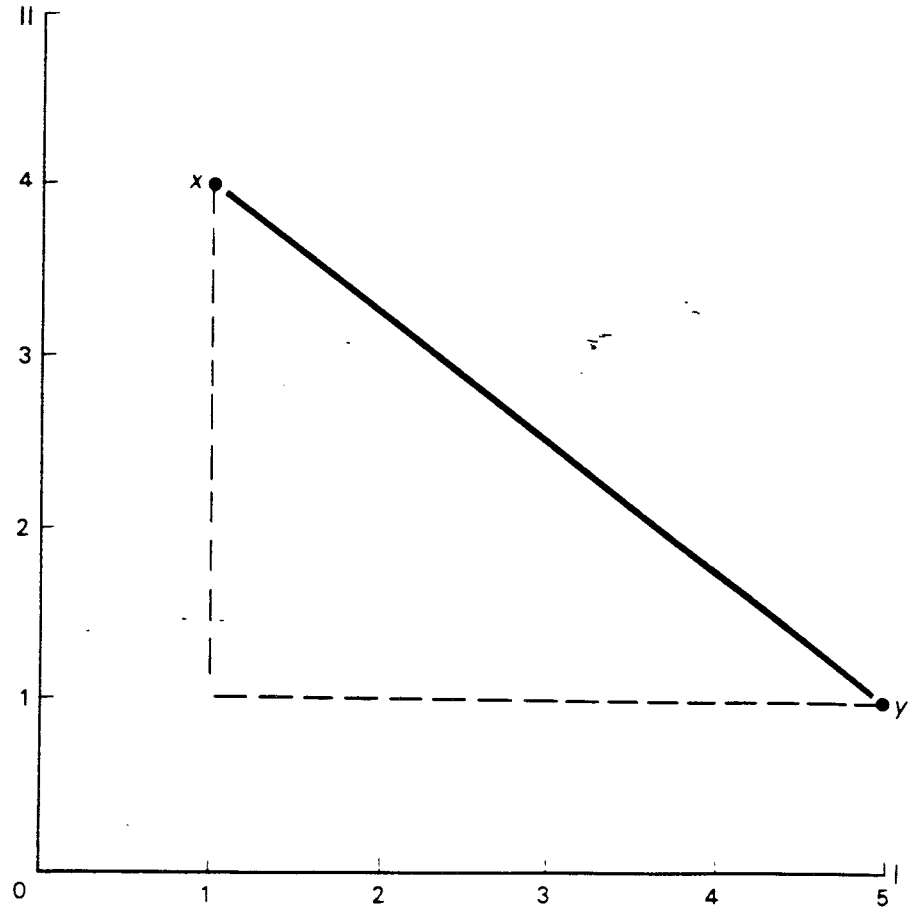
*Carroll and Wish 1974, p. 412 et seq. argue persuasively that the final r-th root may often be usefully ignored, and when this is done a wider range of models qualify as metrics. In the Euclidean case, a number of models are more simply expressed and best understood by treating *squared* distances (i.e. ignoring the final square root). Carroll and Wish (ibid. p. 413) and Shepard (1974, p. 405 et seq.) discuss even more general distance measures, some of which do not even satisfy the triangle inequality.

Euclidean Distance

$$d_{jk} = \sqrt{\sum_a (x_{ja} - x_{ka})^2}$$

where x_{ja} is the co-ordinate of the j th point and x_{ka} is the co-ordinate of the j th point on the a th dimension.

The three commonly-used types of distance mentioned above are illustrated in Figure 5.9. The basic difference lies in the question of whether the differences between objects on each dimension remain separate or merge together ('interact') in producing the overall distance. For $r = 1$ (city block metric) all the dimensional



General Minkowski r–metric

$$d_{xy}^{(r)} = \sqrt[r]{\sum_a |x_a - y_a|^r}$$

City block metric ($r = 1$) (dashed lines)

$$d_{xy}^{(1)} = \sum_a |x_a - y_a| = 4 + 3 = 7$$

Euclidean metric ($r = 2$) (solid line)

$$d_{xy}^{(2)} = \sqrt[2]{\sum_a |x_a - y_a|^2} = \sqrt{4^2 + 3^2} = 5$$

Dominance metric (approximated by $r = 32$)

$$d_{xy}^{(32)} = \sqrt[32]{\sum_a |x_a - y_a|^{32}} = \sqrt[32]{4^{32} + 3^{32}}$$

$$= \sqrt[32]{1.8447_{10}19 + 1.8530_{10}15}$$

$$= \sqrt[32]{1.8449_{10}19}$$

$$= 4.0000125$$

Figure 5.9 *Minkowski metrics*

differences have the same weight in determining the distance; they are simply added together. As r goes to infinity (dominance metric) the largest single difference comes to swamp out all other information. By contrast, the Euclidean distance can be thought of as a compromise where no dimension has a specially important status.

The Euclidean metric is the only one where the orientation of the axes is arbitrary, in the sense that a rotation will leave the distances unchanged. *In all other Minkowski metrics the distances are defined by reference to a fixed set of axes and any rotation will change the distance values.* It is for this reason that axes should be drawn in any configuration where the distance is non-Euclidean.

This property is illustrated by the Minkowski unit-distance (iso-similarity) contour diagram in Figure 5.10. More complex variants are given in Roskam (1968, p. 51) and in Carroll and Wish (1974, p. 417). If all the points at a fixed distance from the origin of a 2-dimensional space are joined, then they form a circle in the case of Euclidean distance (the circle defines the equation $p^2 + q^2 = r^2$, which in this case corresponds to $(x_1 - y_1)^2 + (x_2 - y_2)^2 = d_{xy}^2$ of Figure 5.9). Wherever the dimensions are rotated, the squared dimensional differences still total one, so all are equally permissible. In the case of city block distance, the points at a fixed distance from the origin form a diamond (the diamond is defined by the equation $p + q = r$, corresponding to $(x_1 - y_1) + (x_2 - y_2) = d_{xy}$ of

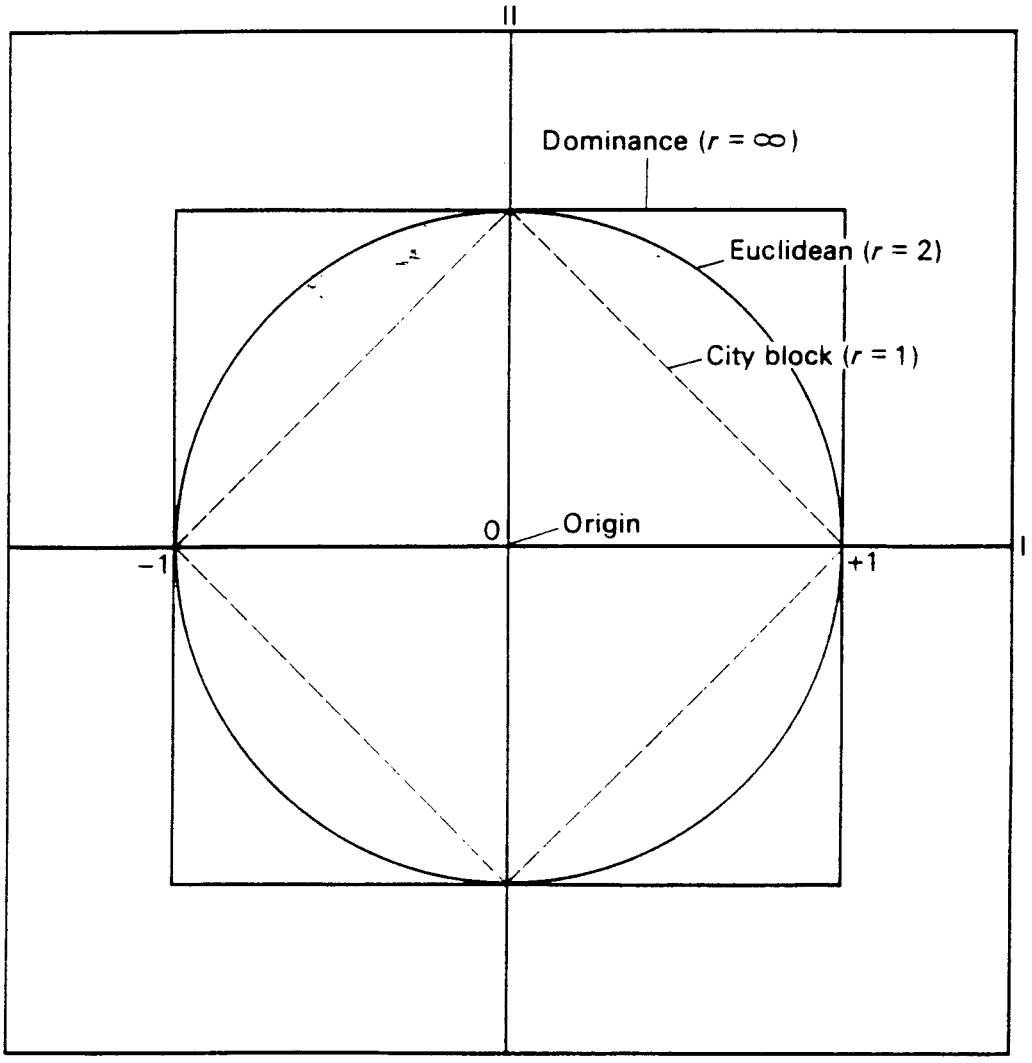


Figure 5.10 *Equal distance contours for 3 Minkowski metrics*

Figure 5.9). If the axes are rotated through anything other than 90° (or multiples of it) the sum of the differences will no longer be the same. For the dominance metric, the unit contour is a cube: until the two differences become equal, only the larger makes any contribution to the distance. Once again, a rotation of anything but multiples of 90° will destroy this relationship.

There is persuasive psychological and empirical evidence (Attneave 1950, Torgerson 1958, Hyman and Well 1968) that the city block metric is particularly appropriate where the characteristics of the objects are obviously compelling or perceptually distinct. By contrast, where the characteristics are more complex the dimensional information begins to merge or blur, and the Euclidean distance will provide a better description. (Compare judging pairs of triangles differing in size and orientation to judging towns in terms of their desirability.) Arnold (1971) has argued that the dominance model provides a better account of data collected by procedures which impose heavy information processing demands on the subject, although the analysis has been questioned by Carroll and Wish (1974).*

A good deal of evidence underlines the conclusion unequivocally argued by Shepard (1974, p. 407) and Carroll and Wish (1974, p. 420) that the Euclidean metric appears to be robust against even extreme departures from its assumptions. Moreover, city block and dominance metrics turn out to be rather subject to local minima and degenerate solutions (information on 8 points can be fit perfectly, in a totally degenerate way, in 3 dimensions, see Shepard 1974). Even if users wish to scale in a 'simpler' metric they are advised to begin with a Euclidean solution and work down (or up) to the preferred metric (Arabie 1973, Shepard 1974).

5.3.3.3 Generalised distance and other metrics

Three other types of distance occur in the MDS(X) programs. The first type (weighted Euclidean distance, a generalisation of Minkowski metrics) is employed in analysing three-way data and is dealt with in the next chapter.

The other two types of distance are simpler than the Minkowski metric and neither are necessarily capable of being represented in a dimensional space. The humblest is simply a metric that obeys the triangle inequality, and the other is the hierarchical clustering or tree-metric that obeys the somewhat more stringent ultra metric inequality.

The simplest type, 'non-dimensional scaling', relaxes not only the additivity requirement of Minkowski spaces, but also the minimum dimensionality of dimensional smallest-space analysis. It does so by dispensing entirely with the idea of a co-ordinate space as embedding the distances, and seeks instead to rescale the data dissimilarities into a set of distances which perfectly obey the triangle inequality, and are as close as possible to being a monotone function of the data. This is achieved by a process of successively increasing the variance of the distances. Intuitively this can be likened to the conformal mapping discussed in 5.2.2, where the largest distances are increased and the smaller ones decreased. This has the effect of forcing down the dimensionality of a space—as in conformal mapping down from a sphere onto a flat plane. In this non-dimensional case, the

*Koopman and Cooper (1974) rightly stress that in two dimensions it is impossible to tell *mathematically* whether the city block or dominance metric is appropriate, since the one is a mathematical transformation of the other—indicated by the fact that the unit contours simply represent a 45° rotation of each other.

analogy does not hold exactly, but it produces 'better behaved' distances. This process is known as maximum variance non-dimensional scaling (Cunningham and Shepard 1974: implemented in MDS(X) as MVNDS) and described in 6.1.7.

The chief virtue of the model is its generality and simplicity: all the other models are special, more restrictive versions of it and only minimal assumptions have to be made to obtain a solution. By dispensing with the assumption of an underlying continuous space, it may also be possible to find a better, more law-like relationship between the original and the rescaled data. Moreover, for the cautious user, this procedure could be used as the first part of the scaling process: obtain a good estimate of the shape of the monotone function without assuming any particular Minkowski metric, and the resulting distances can then be used as input for a more restrictive distance model of one's choice—thus avoiding the dangers of degeneracy and local minima to which non-Euclidean distance scaling is prone. Alternatively, the user might decide to represent the rescaled data in some other way: as a graph, or a tree (i.e. as input to a clustering program).

The other type, the tree-metric, defined by the ultra-metric inequality, was encountered earlier in section 4.3.3.1 as the defining characteristic of a hierarchical clustering scheme (HICLUS program). If a set of data obeys this criterion it can be represented as a dendogram (or rooted tree) where the distance between any two points is defined as the level at which they join (see Figure 4.2).

APPENDIX A5.1 KAPPA AND RELATED MEASURES OF DISCONTINUITY

All Shepard-Carroll (1966) kappa-based measures of continuity have the basic form:

$$\text{kappa} = \left[\underset{\text{(i)}}{\begin{pmatrix}\text{smoothing}\\ \text{ratio}\end{pmatrix}} \times \underset{\text{(ii)}}{\begin{pmatrix}\text{local proximity}\\ \text{weight}\end{pmatrix}} \Big/ \underset{\text{(iii)}}{\begin{pmatrix}\text{normalising}\\ \text{factor}\end{pmatrix}}\right]$$

(i) *Smoothing*

The basic notion of a 'smooth transformation' consists in comparing two uni-dimensional continua, **x** and **y**, in terms of a mapping or transformation which ensures that the intervals or differences between adjacent points (i, j) in the one are of approximately the same size as those in the other, i.e. that the interval $(y_i - y_j)$ is of about the same size (apart from differences in scale) as the interval $(x_i - x_j)$. This is achieved by studying the *ratio* of the differences for each adjacent pair, squaring the result for purposes of convenience. Thus,

$$\left(\frac{\Delta y}{\Delta x}\right)^2 = \left(\frac{y_i - y_j}{x_i - x_j}\right)^2 \tag{1}$$

In the case of MDS, where the data 'distance' (**y**) are being compared to the solution distances (**x**), the differences in (1) become distances, and a simple overall measure of discontinuity or 'lack of smoothness' between the data and the solution is formed by summing the ratio over all $p(p-1)/2$ pairwise data points:

$$\sum_{i \neq j}\sum \left(\frac{\delta_{ij}^2}{d_{ij}^2}\right) \tag{2}$$

(ii) *Local proximity weight*

Local proximity weights w_{ij} are the extension to the multidimensional case of the restriction to adjacent pairs (representing changes in value) in the uni-dimensional case. Shepard and Carroll (1966, p. 582) show that only weights having the form

$$w_{ij} = d_{ij}^{s}, \quad \text{with} \quad s < 0$$

can ensure both that local monotonicity is enforced ($s < 0$), and that the ratio of any two weights remains invariant under change of scale, since solutions are unique only up to similarity transforms. As they indicate, simplicity and experience show that $s = -2$ is a sensible choice, yielding weights of the form: $(1/d_{ij}^2)$, which when multiplied into the discontinuity ratio (1), yields the basic measure of discontinuity,

$$\text{raw kappa} = \sum_{i=j}\sum \frac{(\delta_{ij}^2)}{d_{ij}^4} \tag{3}$$

(iii) *Normalising factor*

Shepard and Carroll (1966, p. 582) define a normalising factor which ensures that kappa reaches a minimum when the solution distances d_{ij}^2 are proportional to the data distances δ_{ij}^2 except for a similarity transform. The simplest such factor is

$$\sum_{i=j}\sum (1/d_{ij}^2)^{-2}. \tag{4}$$

The product of (3) and (4) yields the basic (normalised) kappa index. Gower (1979, p. 3) shows that this normalised kappa measure can be written in a particularly simple and interpretable form:

$$\text{normalised kappa} = \sum_{i=j} w_{ij}\left(\frac{1}{\delta_{ij}^2} - \frac{1}{d_{ij}^2}\right) \tag{5}$$

If the d_{ij} and δ_{ij} are of approximately the same order of magnitude, the weighting factor is approximately equal to $(1/d_{ij}^6)$—which emphasizes fairly starkly how drastic a weighting function it is, giving long distances virtually no influence in determining the final configuration, and giving short (proximate) distances enormous weight. Even small differences in short distances will have very considerable effect on the size of the kappa measure: many users may prefer a less punitive weighting factor.

A5.1.1 Generalised forms of continuity index

Normalised kappa is a special case of the family of continuity indices referred to as 'kappa star'

$$\kappa^* = \sum_{i=j}\sum \frac{(\delta_{ij}^2)^a}{(d_{ij}^2)^b} \Big/ \left[\sum_{i=j}\sum (d_{ij}^2)^c\right]^{-b\,c} \tag{6}$$

(Normalised kappa is the case where $a = 1, b = 2$ and $c = -1$.) If the normalising factor is to keep the kappa index invariant under a similarity transform on the solution space, then the exponents must satisfy the condition $b + c - a = 0$, and c should be negative. (The exponent values can be varied within the PARAMAP

program, where the default values produce the normalised kappa index.)

In terms of the components of the index, a and b affect the continuity ratio, b (and, more indirectly, a) affects the strength of the local monotonicity weight, and b and c affect the normalising factor. Kruskal and Carroll (1969) have argued for $a = b = 1$, thus minimising the local monotonicity weighting and making all 'changes' and distances of equal importance in the minimisation process. They also make the case for reducing the size of the exponents, suggesting two further possibilities:

(i) $a = \frac{1}{2}$, $b = 1$, when the ratio takes the especially simple form of (δ_{ij}/d_{ij}^2), which still preserves local monotonicity weighting, but not in a way that so severely reduces the effect of larger distances: and

(ii) $a = b = \frac{1}{2}$, which removes the local monotonicity weighting and concentrates the effect on the simple ratio of the two distances (δ_{ij}/d_{ij}).

In general, it is necessary that $b > a$ if local monotonicity is to be maintained: the greater the inequality, the more severely discrepancies in representing local structure are penalised, and the less the balancing effect of more distant points.

APPENDIX A5.2 CONVERTING DISTANCE INTO SCALAR PRODUCTS AND BASIC CLASSICAL SCALING

A5.2.1 Conversion of distances into scalar products

In Appendix A2.1 it is shown how to convert scalar products into Euclidean distances. The reverse is often more useful and necessary—how to turn distances into scalar products. This forms the initial stage of most classic metric scaling procedures and is also often used to produce an initial configuration in non-metric models.

(i) *Converting distances into scalar products**

We assume that the distances are genuine and not relative or 'errorful' distances, and to simplify matters we shall assume we are dealing with *squared* distances. Then the required conversion formula is as follows:

$$b_{jk} = -\tfrac{1}{2}(d_{jk}^2 - d_{.j}^2 - d_{k.}^2 + d_{..}^2) \tag{1}$$

where b_{jk} is the scalar product between vectors j and k.

$$d_{.j}^2 = \sum_k^n d_{jk}^2/n, \; d_{k.} = \sum^n d_{jk}^2/n \quad \text{and} \quad d_{..}^2 = \sum_j \sum_k d_{jk}^2/n$$

and n is the number of distances.

Formula (1) can be derived easily from the definition of Euclidean distance so

*This section relies on Carroll (1973). Alternative derivations will be found in Torgerson (1958, p. 255 et seq.).

long as we are dealing with genuine distances (rather than relative or 'errorful' ones) and if we simplify the arithmetic by placing the origin of the space at the centroid of the points, and deal with *squared* distances rather than the distances themselves.

By definition:

$$d_{jk}^2 = \sum_a (x_{ja} - x_{ka})^2 \tag{2}$$

which when multiplied out gives

$$d_{jk}^2 = \sum_a (x_{ja}^2 - 2x_{ja}x_{ka} + x_{ka}^2)$$

$$= \sum_a x_{ja}^2 - 2\sum x_{ja}x_{ka} + \sum x_{ka}^2 \tag{3}$$

This first and third terms on the right are the squared norms of j and k respectively (the vector drawn from the origin to the points concerned), denoted l_j and l_k, hence:

$$d_{jk}^2 = l_j^2 + l_k^2 - 2\sum x_{ja}x_{ka} \tag{3a}$$

The cross product term on the right of (3a) corresponds to the scalar product between vector j and k, denoted b_{jk}, so the last equation can be simplified and rewritten as

$$d_{jk}^2 = l_j^2 + l_k^2 - 2b_{jk} \tag{4}$$

Since l^2 is defined as the averaged squared distance from the origin (i.e. $\sum_j l_j^2/n$) then the following equalities can be shown to hold:

$$d_{\cdot k}^2 = l_\cdot^2 + l_k^2; \quad d_{j\cdot}^2 = l_j^2 + l_\cdot^2 \quad \text{and} \quad d_{\cdot\cdot}^2 = 2l_\cdot^2$$

(where the dot signifies the average over the relevant subscript). Substitution in (4) yields

$$d_{jk}^2 = d_{\cdot k}^2 - d_{j\cdot}^2 + d_{\cdot\cdot}^2 = -2b_{jk}$$

which can be re-arranged as

$$b_{jk} = -\tfrac{1}{2}(d_{jk}^2 - d_{\cdot k}^2 - d_{j\cdot}^2 + d_{\cdot\cdot}^2) \qquad (5) = (1)$$

thus yielding the necessary conversion formula.

As an example, let us return to the distance matrix used in Figure A2.1:

$$\mathbf{D}^2 = \begin{bmatrix} 0 & 5 & 25 \\ 5 & 0 & 8 \\ 25 & 8 & 0 \end{bmatrix}$$

Let us calculate the scalar product b_{32} by (5):

$$b_{32} = -\tfrac{1}{2}(d_{32}^2 - d_{\cdot 2}^2 - d_{3\cdot}^2 + d_{\cdot\cdot}^2)$$

$$= -\frac{1}{2}\left(8 - \frac{13}{3} - \frac{33}{3} + \frac{76}{9}\right)$$

$$= -\tfrac{1}{2}(1.11) = -0.56$$

which is precisely the scalar product (calculated from the centroid as deviate scores) b_{32} given in Appendix A2.1.2.

The conversion formula as can be seen in (5) involves 'double-centring' the squared distance matrix, i.e. removing the row effects, the column effects and adding back in the grand mean.

A5.2.2 The scalar products matrix B and classic scaling

The scalar product matrix, **B**, has a number of properties which are crucial to recovering the space which generated the original distance. Young and Householder (1941) showed that:

(i) If **B** is positive semi-definite (Gramian)—as is necessarily the case if we are dealing with real distances—then by definition its latent roots will be non-negative. This means that the distances can be represented in a real Euclidean space.

(ii) The rank of **B** is equal to the number of dimensions necessary to represent the distances.

(iii) **B** can be factored by conventional methods to obtain a matrix **A**:

$$\mathbf{B} = \mathbf{A}\mathbf{A}'$$

where **A** is a matrix whose elements (a_{ij}) give the projection or co-ordinates of stimulus i on the j th dimension. (These co-ordinates are only unique up to a similarity transform).

Moreover, Eckart and Young (1936) show that if one wishes to obtain a solution in as small a dimensionality as possible (i.e. to approximate a full solution of r dimensions by one in $q \ll r$ dimensions), then the corresponding matrix of co-ordinates (call it **C**, of order q) which minimises the sum of squares of the difference between the full and the approximate solution is given by

$$\mathbf{C} = \mathbf{A}^* \Lambda \mathbf{A}^{*\prime}$$

where Λ consists of the first q latent roots of **B** (in order of magnitude), and **A*** (an incomplete version of **A**) consists of the corresponding q columns or latent vectors of **A**. (see Torgerson 1958, p. 255 et seq. and van de Geer 1971, p. 70 et seq.).

These theorems of Young, Householder and Eckart provide a straightforward way to recover the space that generated a set of distances and produce a close-fitting approximation in a lower dimensionality. The first is achieved by turning distances into scalar products by applying formula (5) and then factoring the resulting matrix to obtain the co-ordinates, which will be unique up to a similarity transformation, and the second is achieved by restricting attention to the first q latent vectors of the matrix.

But these procedures only hold if the data are genuine distances; if they are only 'distance estimates' or relative distances, then we shall encounter the additive constant problem discussed in 5.2.3.2.1 above. Nonetheless, this classic scaling solution turns out to be remarkably robust, and forms an integral part of obtaining the initial configuration for non-metric models, of the now more sophisticated two-way distance metric scaling models and in the basic three-way model, INDSCAL.