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

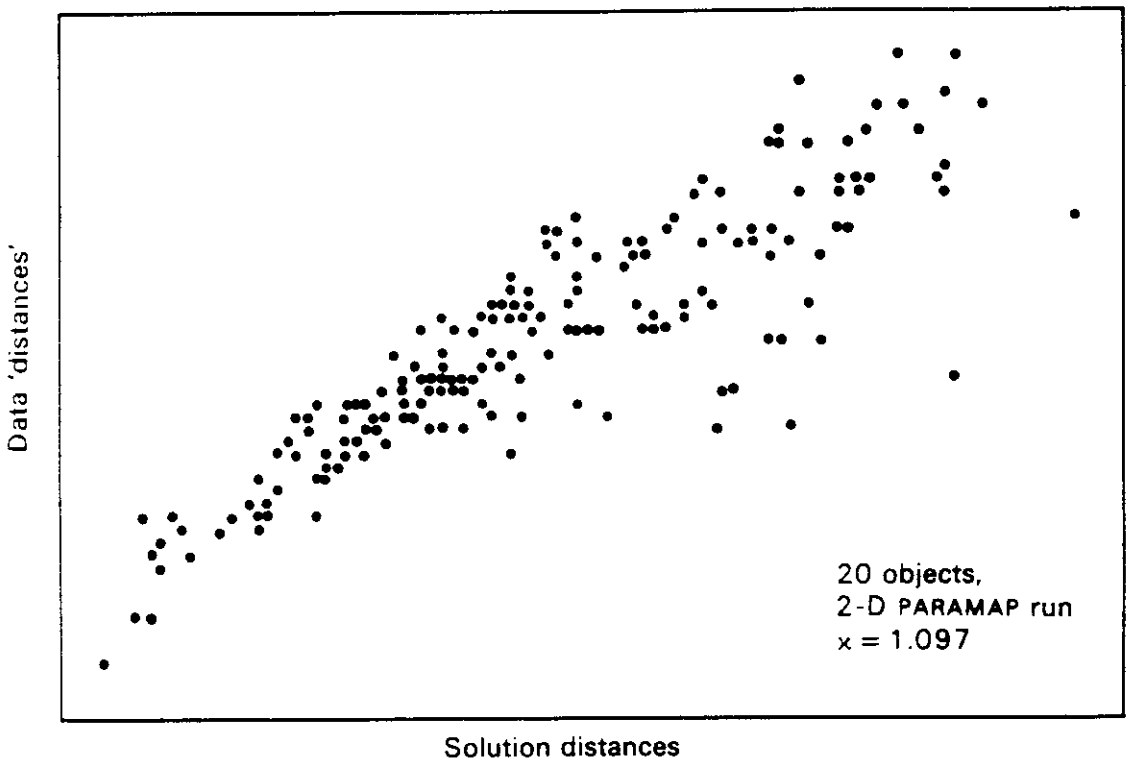


Figure 5.3 *Shepard diagram from continuity scaling*

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

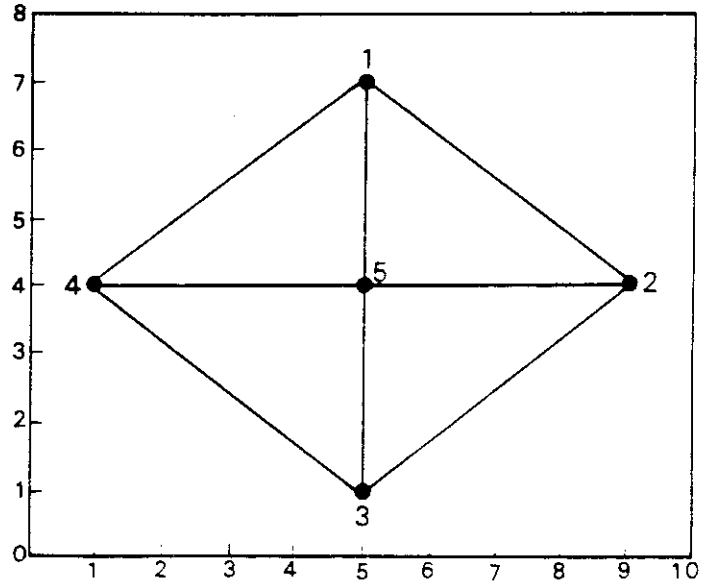


Figure 5.4 *2-dimensional representations of data in Table 5.2b*

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).

5.3.3.2 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.*

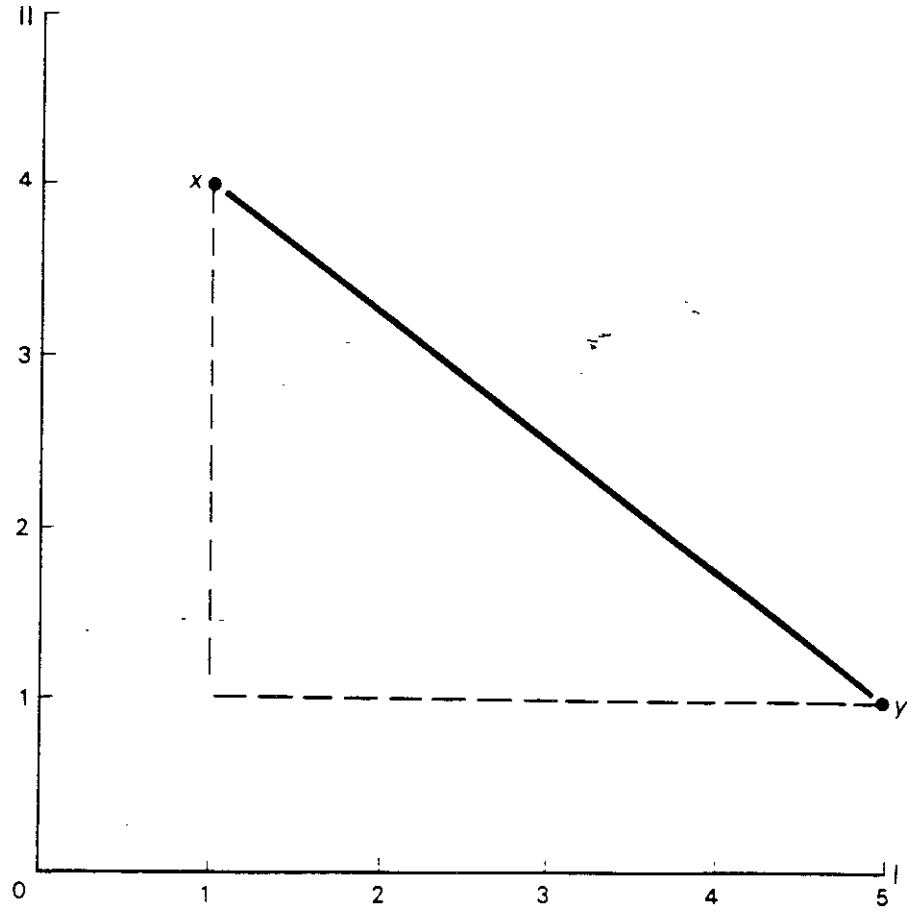
*Carroll and Wish 1974, p. 412 et seq. argue persuasively that the final r-th root may often be usefully ignored, and when this is done a wider range of models qualify as metrics. In the Euclidean case, a number of models are more simply expressed and best understood by treating *squared* distances (i.e. 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

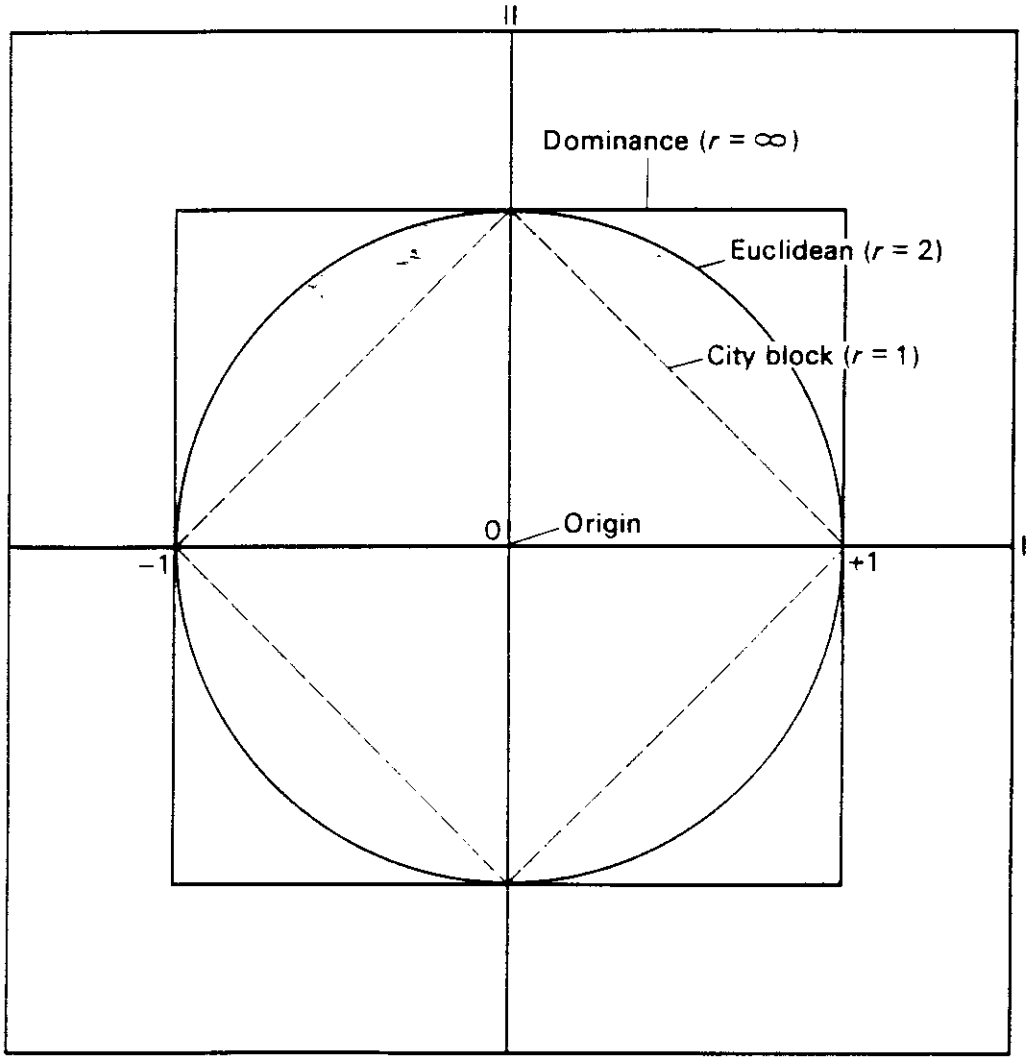


Figure 5.10 *Equal distance contours for 3 Minkowski metrics*

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).

6.1.4 *The basic metric model* (MRSCAL)

Concisely: MRSCAL (MetRic SCALing) provides:

internal analysis of two-way data of a lower triangle format of a (dis)similarity measure

by a Minkowski distance function,

using a linear and/or logarithmic transformation of the data.

*In Coxon and Jones 1978a, global stress$_2$ values of over 0.95 were frequently observed for such heterogeneous data sets, with corresponding local stress$_1$ values of around 0.20. In one case, 169 triadic comparisons of 13 occupations made by a set of policemen produced a global stress$_2$ value of 0.960 when scaled, and 75 out of the 78 global $\hat{d}$ values had the same value!

As we have seen, the assumption that data dissimilarities are a linear function of the distances of the solution historically precedes the monotonic assumption, and the earliest computational techniques for distance model scaling all assumed a linear (or ratio) transformation. Nowadays it is sensible to begin by scaling one's data by the non-metric model, thus making more defensible assumptions about the level of measurement of one's data. But since regular functions are special cases of the general monotonic family of transformations, the Shepard diagram obtained from non-metric scaling should be inspected to determine whether a more regular relationship is discernible between the data and the solution. If so, it makes eminent sense to go on to submit the data to a program such as MRSCAL, which implements the more regular linear and power transformations described in section 5.2.3. The examples in Chapter 3 provide illuminating illustrations of Shepard diagrams where a more regular relationship is evident. For the Scottish mileage data, the Shepard diagram (Figure 3.4) suggests an S-shaped or sigmoid power function*, although in the main range of distances (0.05 to 2.00) the relationship to the data is linear. In the small illustrative example of the similarity of eight crimes (Figure 3.2) the perfect strong monotone function is very close to being linear ($r = 0.97$). By contrast, in the case of the 'real data' example of scaling the co-occurrence frequency of occupational titles, the final Shepard diagram at iteration 23 (Figure 3.14b) shows a distinctly J-shaped (downwardly concave) form, again suggesting the use of a power function (in this case, a negatively accelerated exponential decay function).

As a matter of interest, the 2-dimensional configuration and its associated Shepard diagram obtained from a *linear* scaling of the seriousness of offences data are presented in Figures 6.1 and 6.2. Because the data are very well fit, the equation of the line relating the data to the solution distances is of interest, and attention focuses chiefly upon the slope, as in any other case of linear regression. The linear scaling transformation can be interpreted as indicating that, if the data are increased throughout by an additive constant of 0.251, then the distance between the points will on average be one quarter (0.241) of the difference between those saying they are 'unalike in their seriousness'. The advantage of a regular transformation function, then, is that it is possible to extrapolate beyond the original data (and in this sense 'predict' further data) if the model is correct. (In this particular example, we should not expect the transformation to be linear throughout its range, since there is an upper limit of 1.00 on the data values.)

The MRSCAL program in the MDS(X) series implements the linear and power models outlined in Roskam (1972) and contains options for implementing the city block and Euclidean distance metrics, among others.

6.1.5 Parametric mapping (PARAMAP)

Concisely: PARAMAP (PARAmetric MAPping) provides:

internal analysis of either a rectangular or a square symmetric two-way data matrix

by a distance model which maximises continuity or local monotonicity.

*In fact, a logistic function would best fit these data, but an ordinary logarithmic transform closely approximates its form.

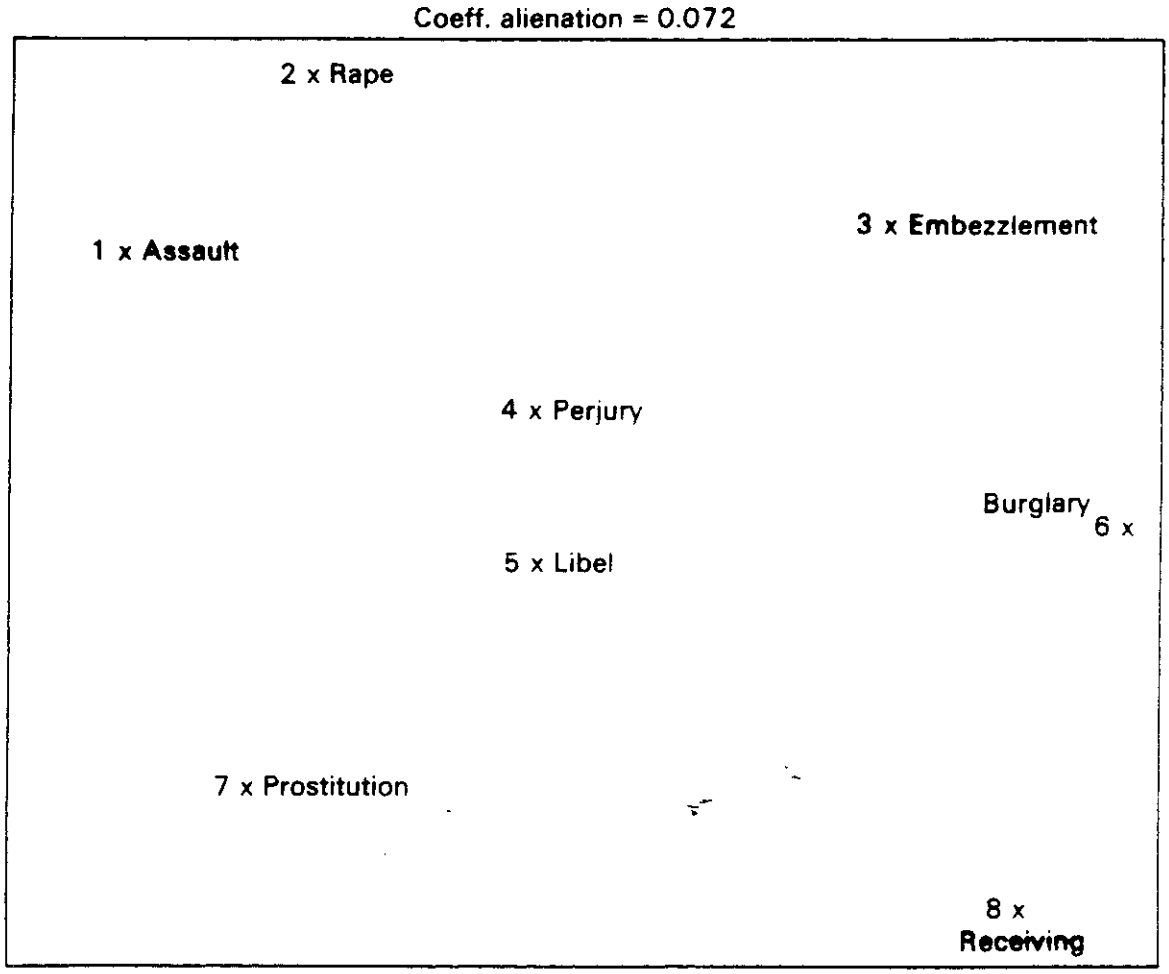


Figure 6.1 *MRSCAL 2-D solutions to data of Table 3.1*

The PARAMAP program accepts data either in the form of symmetric lower-triangular distances *or* in rectangular form consisting of profile values or spatial co-ordinates, but it only represents the objects (rows) in the latter case. The 'smoothness' or continuity transformation used in PARAMAP is the kappa family of continuity indices described in detail in 5.2.2.1 and in Appendix A5.1. The default values of the program parameters produce the simple 'normalised kappa' index (Appendix A5.1, equation 5), and the effect of these and other variants on the representation of the data are discussed in some detail at that point.

The main distinguishing characteristics of parametric mapping are:

(i) the faithful preservation of the *local* information (small distances) around each point, virtually ignoring large distances (the user is thus given control of the degree of local monotonicity); and

(ii) the 'flattening' of configurations into as small a dimensionality as possible, by increasing the size or variance of the largest distances.

The two are related: the price paid for preserving local information in a low dimensionality is the considerable distortion of global information. This fact should be borne in mind when interpreting PARAMAP solutions, since it contradicts the usual MDS maxim that configurations are globally stable, in the sense that the main features are reliably fixed but are locally unstable in that the points in a configuration can be moved around slightly without any major change in stress.

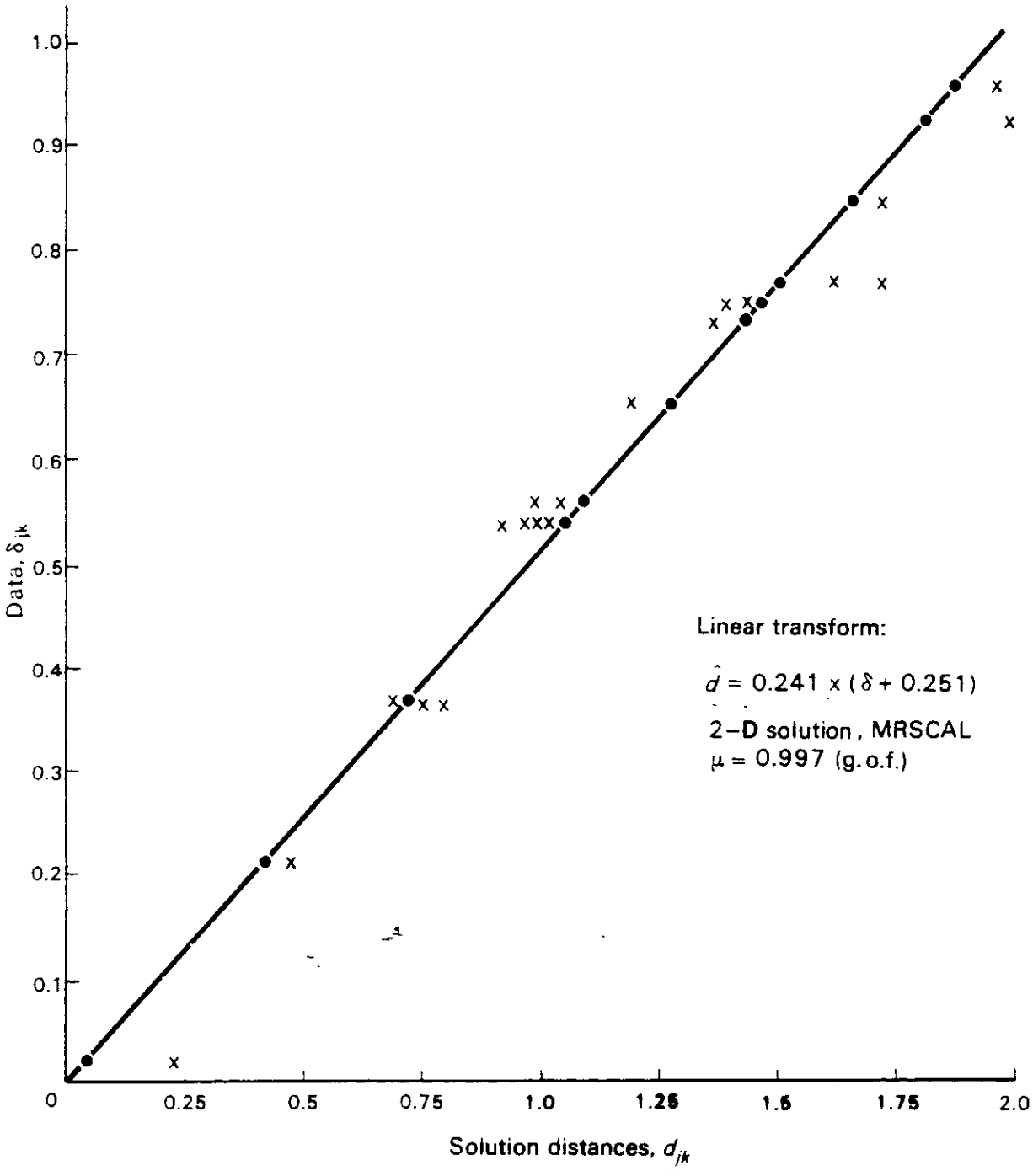


Figure 6.2 *Shepard diagram metric (linear) solutions to data of Table 3.1*

Moreover, because of the considerable non-linearity of the continuity rescaling transformation, any external properties the user wishes to represent should be mapped into a configuration using the *non-linear* option in PROFIT, which optimises an index akin to kappa. If hierarchical clustering is used to interpret a PARAMAP solution, only the initial stages of the clustering should be mapped into the configuration, because the largest distances are bound to be badly represented as a result of the continuity criterion.

The most dramatic examples of parametric mapping, which show these properties most obviously, occur for the mathematical structures of points defined at regular intervals on the circle (mapped into the line), on the sphere and on the torus, when mapped down into two dimensions (see Shepard and Carroll 1966,

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:

$$b_{jk} = -\tfrac{1}{2}(d^2_{jk} - d^2_{.j} - d^2_{k.} + d^2_{..}) \qquad (1)$$

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

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

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.

APPENDIX A3.1 COMPARISON OF MEASURES OF FIT BETWEEN DATA AND SOLUTION USED IN NON-METRIC MDS

1 All measures of fit used in the basic model compare a set of disparities (ratio-level quantities which are a function of the data, $d^0_{jk} = f(\delta_{jk})$, which are *monotonic* for ordinal data and *linear* for metric or interval data with the ratio-level distances, d_{jk}. Two forms of comparison are used:

(i) *the difference* $(d_{jk} - d^0_{jk})$, which forms the basis of badness-of-fit measures, since the greater the discrepancy between a solution and the data, the greater will be the differences; and

(ii) *the scalar product* $(d_{jk}d^0_{jk})$, which forms the basis of goodness-of-fit measures, since the greater the covariance (or the less the angular separation) between data and the solution, the greater will be the scalar products.

2 The basic measure of goodness-of-fit used in non-metric programs emanating from the Bell Laboratories is *stress*, and in particular (normalised) stress$_1$, based upon Kruskal's BFMF disparities, $\hat{d}_{jk}$. These and alternative measures are dealt with in sections 3.3 and 3.5.2 and are extensively discussed in Kruskal and Carroll, 1969.

3 The measures used by Lingoes (Michigan), Guttman (Israel) and Roskam (Nijmegen) include stress measures (often based upon Guttman's rank-image disparities, d^*_{jk}), but also include a number of less familiar measures. In particular:

(a) Mu $(\mu) = \sum d_{jk}d^0_{jk}$

$$\sqrt{\left\{\sum d^2_{jk} \sum \left(d^0_{jk}\right)^2\right\}}$$

(Goodness of fit, varies between -1 and $+1$)

This measure is akin to the Pearsonian correlation coefficient. It is independent of the scale of both the distances *and* the disparities if the data are scaled by a ratio transformation, as in the metric MDS model (with the logarithmic option) implemented as MRSCAL in the MDS(X) series.

If the data are scaled by a linear transformation, as in MRSCAL (under the linear option), then mu is formally identical to the Pearsonian correlation coefficient r. It may also be used for monotone transformations, but it is not entirely clear whether it is dependent in this case on the scale of the disparities.

(b) Alienation (K) $= \sqrt{(1 - \mu^2)}$ (Badness of fit, varies between 0 and 1)

This measure is akin to stress$_1$ and in some cases is identical to it. In any event, K is strictly monotonic with stress$_1$. The coefficient measures the extent of residual variance from the fitted regression.

(c) (Normalised) Phi (ϕ) = (Raw Stress/(2 × NF 1))
$= \sum (d_{jk} - d^0_{jk})/2 \sum d^2_{jk}$
(Badness of fit, varies between 0 and 1)

This measure is also akin to stress_1, but differs in the scaling factor—twice that of stress_1—and in the fact that the index is not reduced by its square root. It differs from the coefficient of alienation K, in being based upon the difference, rather than the scalar product, of the distances and the disparities.

Strictly speaking, any of these three measures may be used either with Kruskal's monotone regression disparities $\hat{d}_{jk}$, or Guttman's rank images d^*_{jk}, although by convention they are normally used with the latter.

4 *Relation between fit measures*

Relationships between the fit measures depend most importantly on whether Kruskal's $\hat{d}$ or Guttman's d^* quantities are being used. (In reporting measures of fit, users should always indicate which fitting quantities are being referred to and MDS(X) programs indicate the referent quantities as d-hat and d-star respectively.) In general, for any of these badness-of-fit measures, a measure based on d^* will be higher (indicating worse fit) than the same measure based on $\hat{d}$, since the former attempts to preserve strong monotonicity and the latter preserves only weak monotonicity with the data.

This can best be exemplified by relating various measures to μ, which represents the cosine of the angle separating the distances and the fitting quantities, d^0, in the measurement space (see Roskam 1969, p. 13):

	Fitting quantities (disparities)	
Measure	Kruskal's $\hat{d}$ (weak)	Guttman's d^* (strong)
μ	$\cos(d, \hat{d})$	$\cos(d, d^*)$
Stress_1 (S_1)	$\sqrt{(1 - \mu^2)}$	$\sqrt{2(1 - \mu)}$
Phi	$\frac{1}{2}(1 - \mu^2)$	$(1 - \mu)$
K (Alienation)	$(= S_1)$	$\sqrt{(1 - \mu^2)}$

Other useful relationships are as follows:

Alienation and phi (i) $K = \sqrt{1 - (1 - \phi)^2} = \sqrt{(1 - \mu^2)}$

Alienation and stress_1 (ii) If $\hat{d}$ is used, $K = S_1$, and if d^* is used, $K = S_1\sqrt{1 - (\frac{1}{2}S_1)^2}$

Phi and stress_1 (iii) $\phi = \frac{1}{2}S^2$