Appendix A

Multidimensional Scaling

A.1 INTRODUCTION

A problem encountered in many disciplines is how to measure and interpret the relationships between objects (Aarts [4]). A second problem is the general lack of a mathematical relationship between the perceived response and the actual physical measure. Sometimes relationships are rather vague. How much does the character of one person resemble that of another? Or in the case of this book, to what extent are various processing methods alike? How do we measure and what scale do we need? In the following text, we discuss some scales and techniques and give some examples.

A short but authoritative introduction to multidimensional scaling(MDS) is Kruskal's book [151]. A comprehensive survey of the development of MDS is that of Caroll & Arabie [46], which cites 334 references, mostly published during the 1970s. More recent are the surveys by Young [303] and (on general scaling) by Gescheider [88]. The latter is more about sensory and cognitive factors that affect psychophysical behavior than about measurement and computational aspects. A review intended for a wide, generally scientific audience, concerning the models and applications of MDS and cluster analysis, has been provided by Shepard [246].

A.2 SCALING

The purpose of scaling is to quantify qualitative data. Scaling procedures attempt to do this by using rules that assign numbers to qualities of things or events. Multidimensional scaling is an extension of univariate scaling (App. A.7); it is simply a useful mathematical tool that enables us to represent the similarities of objects spatially as in a map.

MDS models may be either metric or non-metric, depending on the scale of measurement used in collecting the data. For metric scaling, the collected data should be measured using an interval or ratio scale. In the former case, the unit of the 'yardstick', used for measuring the phenomenon, as well as the zero point (offset) are unknown. For a ratio scale, the zero point is known but there is an unknown scaling factor. For non-metric MDS, only the ranking order of the data values is used; the data are, or are used at, an

Audio Bandwidth Extension: Application of Psychoacoustics, Signal Processing and Loudspeaker Design. E. Larsen and R. M. Aarts
 ISBN 0-470-85864-8

ordinal level. However, it is sometimes possible to recover metric distances obtained by non-metric MDS, as will be shown in an example later on.

In order to obtain a spatial map from an MDS computer program, we only need to apply a set of numbers as the input. To all (or most) combinations of pairs out of a group of objects, a number is assigned, which expresses the similarity between the objects of that pair. Such numbers are sometimes referred to as proximities. MDS procedures will then represent objects judged similar to one another as points close to each other in the resulting spatial map. Objects judged to be dissimilar are represented as points distant from each other.

MDS programs that use direct similarity measures as input have the advantage of being low in experimental contamination. They do not require a prior knowledge of the attributes of the stimuli to be scaled.

A.3 EXAMPLE

An obvious procedure for obtaining similarity data is to ask people directly to judge the 'psychological distance' of the stimulus objects. Another way is the method of triadic comparisons (Levelt *et al.* [159]). This has the advantage that it simplifies the subject's task, because only the ranking order of three presented stimuli is asked for. However, there can be some drawbacks, as pointed out by Roskam [229]. A practical problem arises when for a complete experiment the number of triads (all possible combinations of three stimuli out of the set of all stimuli) is considered too large. It can be reduced by using an incomplete balanced block design (BIBD).

As an example of MDS using triadic comparisons, consider the following. Suppose someone with a good topographical knowledge of The Netherlands is asked to give the nearest city and the most distant city out of three given cities. The same question is asked for three other cities, and so on, until each distance from one city to another (each out of a total list of 14 cities) is considered. A matrix M can be constructed so that for the three cities (i,j,k) the two closest together, for example, (i,j) contribute 0 points to the matrix element M(i,j), the next closest pair, for example, (j,k), adds 1 point, and the remaining pair adds two points to M(i,k). The (dissimilarity) matrix obtained in this way resembles an ordinary distance table. If the phrases most distant and nearest in the question are interchanged, one obtains a similarity (data) matrix.

Instead of relying on a topographer, we used an ordinary distance table as input for the program. The program we applied was KYST-2a, pronounced 'kissed', formed from the names Kruskal, Young, Shepard, and Torgerson. It gives the coordinates of the cities in one or more dimensions. The analysis was carried out for both the metric case (with linear regression) and the non-metric case. In the latter case, the actual number of miles was not used. However, the ranking order of the calculated interpoint distances should be, as far as possible, the same as the ranking order of the interpoint distances in the given distance matrix. The results of both the metric and the non-metric cases were practically the same. Only the results of the latter case will be discussed in the following.

All calculations were carried out in the Euclidian space (Minkowski's parameter = 2). A measure of the goodness of fit between both rankings is called stress, which can to some extent be compared with a least-squares sum in an ordinary fitting procedure. The stress value in this particular case is 0.249 for one dimension, 0.028 for two dimensions,

and 0.013 for three dimensions. It appears that a two-dimensional fit is a good one. The decrease of stress in three dimensions is rather weak, while the deterioration due to a low dimensionality is obvious. The results of the calculations are plotted in Fig. A.1; the solid points are the real locations, whereas the small circles represent the calculated places. As the figure shows, in this particular case it is possible to derive metric data from non-metric input data. The orientation of the map is arbitrary; there is no real North–South axis. For convenience only, the contour of the Netherlands and a compass needle are drawn.

A second example is from Ekman's [63] similarity judgement among 14 colours varying in hue. Subjects made ratings of qualitative similarity for each pair of combinations of colours ranging in wavelength from 434 to 674 nm. Shepard [245] applied a non-metric MDS procedure to the similarity ratings and extracted the underlying structure depicted by Fig. A.2. The underlying structure recovered from Ekman's similarity data was simply

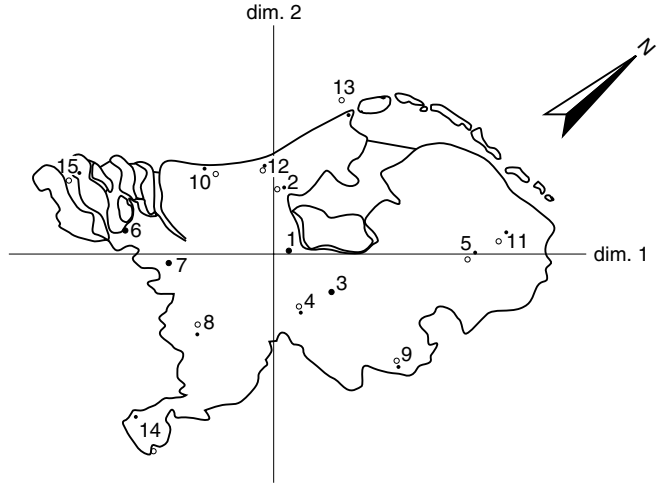

Figure A.1 Configuration with real (solid) and calculated (circles) locations

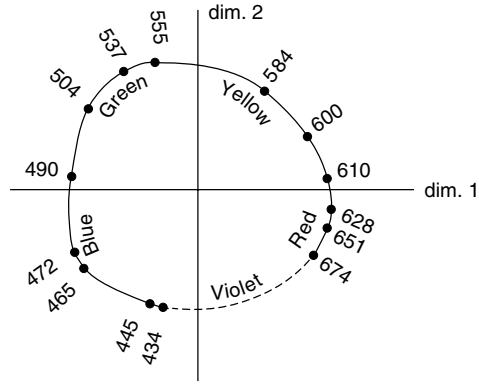

Figure A.2 Colour circle (From Shepard's [245] analysis of Ekman's data)

the conventional colour circle, with the colours arranged along the smooth contour in order of increasing wavelength.

A.4 PROCEDURE

The input data, δ_{ij} or *proximities*, are numbers that indicate how similar or how different two objects are, or are perceived to be. The distance between the points i and j in the configuration, which reflects the 'hidden structure' is denoted by d_{ij}. The basic concept takes the form that

$$f(\delta_{ij}) = d_{ij}, \tag{A.1}$$

where f is of some specified type. The discrepancy between $f(\delta_{ij})$ and d_{ij} is then

$$f(\delta_{ij}) - d_{ij}. \tag{A.2}$$

An objective function that is called stress is

$$S_1 = \sqrt{\frac{\sum_i \sum_j [f(\delta_{ij}) - d_{ij}]^2}{\sum_i \sum_j {d_{ij}}^2}} \tag{A.3}$$

The values $f(\delta_{ij})$ are often called fitted distances and denoted by $\hat{d_{ij}}$; sometimes, they are also called 'disparities'. When the *only* restriction for f is that it has to be monotonous, then the procedure is of the non-metric type.

A.5 PRECAUTIONS CONCERNING THE SOLUTION

The interpretation and generation of the configuration map should both be monitored carefully, as undesirable results can occur, which render any use of the configuration inadvisable (Kruskal *et al.* [152]). *Before* attempting to interpret the configuration, the user should *always* check for these possibilities, beginning with the inspection of $\delta - d$, the Shepard diagram (i.e. the scatter plot of recovered distances versus data values). Some anomalous $\delta - d$ configurations are discussed below. The relation between stress and significance is studied by Wagenaar and Padmos [292], and is discussed in the next section.

Jaggedness of the fitted function: The function relating distances to data values will always be somewhat jagged. However, this function should in fact approximate a smooth and continuous curve. Since the user is assuming an underlying continuous function, a configuration associated with a step function is undesirable. There are, however, two possible remedies for step-functions. One is the possibility of a local minimum, hence different initial values have to be tried. The second remedy is to specify a stronger form of regression.

Clumping of stimulus points: This is when several distinct objects occur at the same position in the Shepard diagram. The phenomenon is associated with undesirable behaviour of the fitted function (i.e. the d values) in the region of the smallest dissimilarities. The antidote for undesirable clumping is to declare a different form of regression, perhaps with a preliminary transformation of the data as well.

Degeneracy: A degenerate solution is an extreme case of both clumping (clustering) and jaggedness. Usually the stress is in this case zero, or nearly so. Hence, a very small stress value can indicate an utterly useless solution instead of an exceptionally good one. This result often occurs when, for one or more subsets of the stimuli, the dissimilarities within that subset are smaller than the dissimilarities between stimuli in that subset and the remaining stimuli (Shepard [245]). Possible solutions are (1) separate the clusters, (2) scale them separately and combine them in a later run (3) use the FIX option, or (4) a form of regression stronger than monotone can be specified.

A.6 SIGNIFICANCE OF STRESS

For the representation of an arbitrary dissimilarity matrix by distances between n points in m dimensions, a probability $p(s)$ exists that a stress value $\leq s$ will be obtained by chance. For the determination of the probability distributions $p(s)$, the dissimilarity matrices contained the numbers from 1 to $0.5n(n-1)$ and were attributed randomly to the cells. In this way, 100 'random scatters' were produced and analyzed by the MDS technique in various dimensions. The results are in Table A.1. The empty cells in Table A.1 correspond to conditions where more than 5% of the scatters have a stress smaller than 0.5%; it is advisable never to use MDS in these conditions.

A.7 UNIVARIATE SCALING

The purpose of scaling is to quantify the qualitative relationships between objects by scaling data. Scaling procedures attempt to do this by using rules that assign numbers to qualities of things or events. Here we discuss univariate scaling, in contrast to multidimensional scaling. Univariate scaling is usually based on the law of comparative judgement

Table A.1 The maximum stress in percentages, which can be accepted at a significance level of $\alpha = 0.05$ for n points in m dimensions, from Table III of Wagenaar and Padmos [292]

	m = 1	2	3	4	5
n = 7	20	7	–	–	–
8	27.5	10	1.5	–	–
9	30.5	13	5.5	1	–
10	34	15	7	3	–
11	35	18	9.5	4.5	1
12	39.5	20.5	10	6.5	3.5

(Torgerson [276], Thurstone [270]). It is a set of equations relating the proportion of times any stimulus i is judged greater, or has higher appreciation for a given attribute, than any other stimulus j. The set of equations is derived from the postulates presented in Torgerson [276]. In brief, these postulates are:

1. Each stimulus when presented to an observer gives rise to a discriminal process, which has some value on the psychological continuum of interest (e.g. in the context of this book, appreciation for a particular processing method as judged by listening to a processed signal).
2. Because of momentary fluctuations in the organism, a given stimulus does not always excite the same discriminal process. This can be considered as noise in the process. It is postulated that the values of the discriminal process are such that the frequency distribution is normal on the psychological continuum.
3. The mean and standard deviation of the distribution associated with a stimulus are taken as its scale value and discriminal dispersion respectively.

Consider the theoretical distributions S_j and S_k of the discriminal process for any two stimuli j and k respectively, as shown in the upper panel of Fig. A.3. Let $\overline{S}_j$ and $\overline{S}_k$ correspond to the scale values of the two stimuli and σ_j and σ_k to their discriminal dispersion caused by noise.

Now we assume that the standard deviations of the distributions are all equal and constant (as in Fig. A.3), and that the correlation between the pairs of discriminal processes is constant; this is called 'Condition C' by Torgerson [276]. Since the distribution of the difference of the normal distributions is also normal, we get

$$\overline{S_k} - \overline{S_j} = cx_{jk}, \tag{A.4}$$

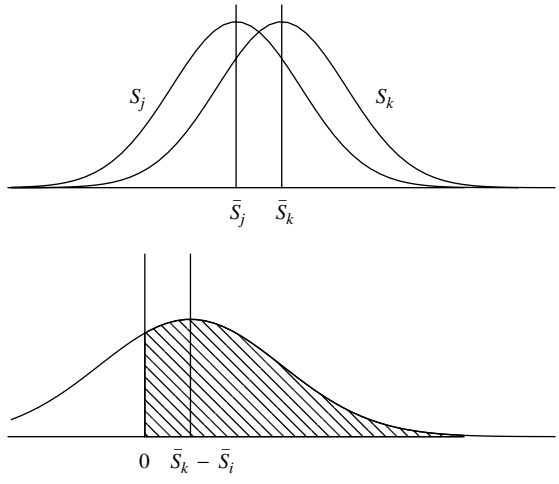

Figure A.3 The upper panel shows two Gaussian distributions corresponding to stimuli j and k, having different mean values ($\overline{S}_j$ and $\overline{S}_k$). The probability that k is judged to be larger than j is given by the shaded area in the lower panel

where c is a constant and x_{jk} is the transformed (see Eqn. A.7) proportion of times stimulus k is more highly appreciated (or judged greater) than stimulus j. Equation A.4 is also known as Thurstone's case V. The distribution of the discriminal differences is plotted in the lower panel of Fig. A.3. Equation A.4 is a set of $n(n-1)$ equations with $n+1$ unknowns, n scale values and c. This can be solved with a least-squares method. Setting $c = 1$, and the origin of the scale to the mean of the estimated scale values, that is,

$$1/n \sum_{j=1}^{n} s_j = 0, \tag{A.5}$$

we get

$$s_k = 1/n \sum_{j=1}^{n} x_{jk}. \tag{A.6}$$

Thus, the least-square solution of the scale values can be obtained simply by averaging the columns of matrix $\mathbf{X}$; however, the elements x_{jk} of $\mathbf{X}$ are not directly available. With paired comparisons we measure the proportion p_{kj} that stimulus k was judged greater than stimulus j. This proportion can be considered as a probability that stimulus k is in fact greater than stimulus j. This probability is equal to the shaded area in Fig A.3, or

$$x_{jk} = \text{erf}(p_{jk}), \tag{A.7}$$

where erf is the error function (Abramowitz and Stegun [12, 7, 26.2]), which can easily be approximated (Abramowitz and Stegun [12, 26.2.23]). A problem may arise if $p_{jk} \approx \pm 1$ since $|x_{jk}|$ can be very large. In this case, one could simply replace x_{jk} by a large value.

It may be noted that this type of transformation is also known as Gaussian transform, where instead of the symbol x, z is used, known as the z scores. Instead of using Eqn. A.7, other models are used, for example, the Bradley–Terry model, see David [56]. All forms of the law of comparative judgement assume that each stimulus has been compared with other stimuli a large number of times. The direct method of obtaining the values of p_{jk} is known as the method of paired comparisons, see, for example, David [56].