CURVED AXES AND TRAJECTORIES FOR MULTIDIMENSIONAL SCALING, WITH APPLICATIONS TO SENSORY AND CONSUMER DATA

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy

by

Stephen John Bennett

August 2008

Acknowledgements

Firstly, I would like to thank Unilever for giving me the opportunity to study for my PhD, especially to Dr Trevor Cox. Without Trevor's support and guidance throughout my PhD, much of the work in this thesis would have been more difficult.

Secondly, I would like to thank Prof. Raj Bhansali and his group at the University of Liverpool.

Finally, my thanks go to my wife, Julie Bennett, for her total support throughout the last seven years.

The work in Chapter 2 was presented in the paper 'Linear and curvi-linear axes in multidimensional scaling plots' at the Royal Statistical Society conference RSS02 in Plymouth in 2002.

The work in Chapter 3 was presented as a poster entitled 'An MDS approach to multivariate paired comparison data' at Sensometrics 2008 at Brock University, Ontario, Canada in 2008. This won the best poster award, as judged on its scientific context by the organising committee.

Finally, the work in Chapter 4 was presented at the International Conference on Methodology of Longitudinal Surveys conference at Colchester in 2006, in a paper entitled 'Dynamic multidimensional scaling - a new approach to analysing longitudinal panel data'.

Contents

Abstract

The analysis of sensory and consumer-derived data involves the use of many different statistical techniques. The vast majority of these are multivariate in nature - for example, multidimensional scaling (MDS) and biplots. However, univariate techniques such as repeated measures analysis of variance and the Bradley-Terry model for paired comparison data are also common.

This thesis introduces enhancements to MDS based on the use of curved axes and trajectories.

Firstly, curved axes representing attributes are overlaid onto MDS maps, in an attempt to describe the maps in more detail. Different functions are used to define the axes, resulting in a biplot-like configuration that enables improved understanding of the data.

Secondly, the method of univariate paired comparisons is extended to the multivariate case, resulting in a methodology for visualising multivariate paired comparison data via an MDS-style approach. Again, the result is a biplot-like configuration, which allows projection of pairs of objects onto axes in order to determine which of the pair is generally preferred or chosen on each attribute.

Finally, dynamic MDS is introduced as method for visualising repeated multivariate data, initially arising from consumer questionnaires carried out

over a period of time. Trajectories are used to show how the proximities of the objects change with time. An extension shows that any continuous variable can be used in place of time.

The methodologies are demonstrated on a series of simulated data sets, and real data from the Home and Personal Care Industry.

Chapter 1

Introduction

1.1 What is Multidimensional Scaling?

Multidimensional Scaling (MDS) is a series of statistical techniques concerned with displaying certain kinds of data spatially using a map. The basic premise is that points on the map represent objects, and the more similar that two objects are to each other in multivariate space, then the closer the two representative points will be together on the map. MDS can thus be used to analyse any data that represents how similar (or dissimilar) objects are to one another. For this reason, MDS has found application in a broad range of disciplines, including physics (Lilensten et al. (2007)), psychology (Yang and Lin (2008)), linguistics (Verheyen et al. (2007)), political science (Hook (2007)), genetics (Wang et al. (2007)), sensory science (Lim and Green (2007) and Yoshioka et al. (2007)), and shape analysis (Axelsson (2007) and Cooke et al. (2007)). In each case, MDS is used to construct a spatial representation of the similarity amongst objects, with the purpose of discovering relationships or patterns. A discussion about these and further examples can be

found in Section 1.8. The development of MDS was largely motivated by a desire for a psychophysical scaling method that did not presuppose a knowledge of the attributes on which stimuli differ (Torgerson (1958), Young and Hamer (1987) and Mead (1992)).

A detailed theory of MDS can be found in many books such as Schiffman et al. (1981), Coxon (1982), Everitt and Rabe-Hesketh (1997), Cox and Cox (2000), and Borg and Groenen (2005), the last of which also looks in depth at some of the standard MDS computer programs which have been developed. In addition, a chapter on MDS is often included in many Multivariate Analysis books such as Everitt and Dunn (2001) and Cox (2005).

1.2 Multidimensional Scaling - an example

The basic concept of MDS can be demonstrated using an example adapted from Kruskal and Wish (1978), and the type of which is used in many textbooks on multivariate analysis. Consider Table 1.1, showing the distances by road between ten English towns and cities. The data within this table can easily be constructed, if a map were available, by using a ruler and measuring the distances involved.

MDS is designed to solve the opposite (more difficult) problem of constructing a map from such data. From Table 1.1, it can be seen that the two closest places are Liverpool and Manchester, which are 54.9km apart. Thus it can be expected that the points representing these two cities will be close together on the map. Additionally, London and Carlisle are the furthest apart (497.3km), and so these two cities are expected to be far apart on the map. By carrying out MDS on these distances, a map of the cities is obtained

Town	Birm	Bris	Car	Leeds	Liv	Lond	Manc	Newc	Nor	Oxf
Birmingham	0	139.9	313.8	190.7	157.2	191.2	135.4	333.6	255.7	107.4
Bristol	139.9	0	444.9	333.8	288.4	190.9	266.7	476.8	384.9	116.9
Carlisle	313.8	444.9	0	188.6	199.4	497.4	193.5	93.9	455.9	429.7
Leeds	190.7	333.8	188.6	0	116.7	314.3	67.9	154.5	279.2	270.4
Liverpool	157.2	288.4	199.4	116.7	0	342.6	54.9	270	344.4	274.4
London	191.2	190.9	497.4	314.3	342.6	0	319.1	457.3	189.8	100.3
Manchester	135.4	266.7	193.5	67.9	54.9	319.1	0	221.4	295.5	251.1
Newcastle	333.6	476.8	93.9	154.5	270	457.3	221.4	0	407.1	412.3
Norwich	255.7	384.9	455.9	279.2	344.4	189.8	295.5	407.1	0	273.1
Oxford	107.4	116.9	429.7	270.4	274.4	100.3	251.1	412.3	273.1	0

Table 1.1: Road distances (km) between selected towns and cities in the UK (www.theaa.com (2008))

as shown in Figure 1.1, which almost perfectly recreates the geographical arrangement of the cities.

A reason for possible discrepancies between the map obtained and the usual geographical map is the fact that the distances in the table are those taken to travel between the cities by road, not the Euclidean distance that the MDS algorithm uses. The map also has an unusual orientation when compared to normal geographical maps - North is to the right hand side of the map, and East is towards the top. This is because the configuration obtained by MDS is not unique - the map is produced solely from the distances, and can be rotated, translated or reflected without changing the distances between the points. If the configuration in Figure 1.1 was rotated 90° anti-clockwise, and then reflected about the vertical axis, the 'expected' orientation would be obtained, and this still would be a valid MDS solution. Section 1.6 deals with

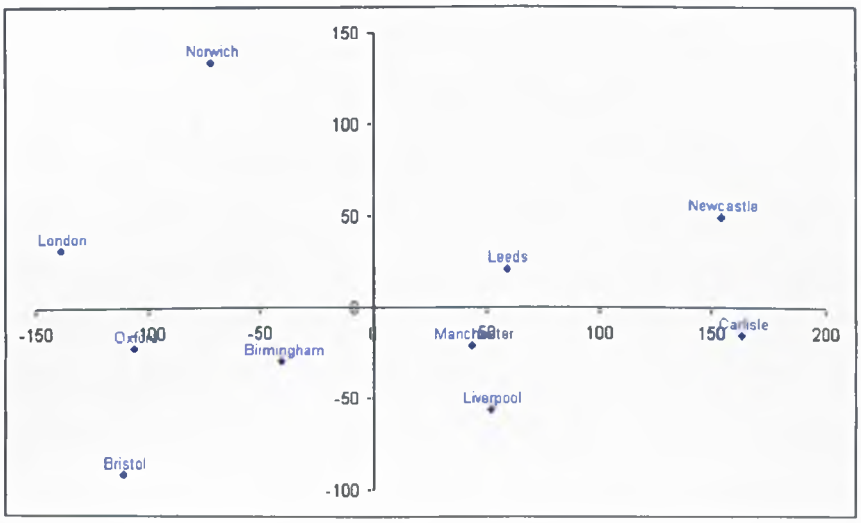

Figure 1.1: Metric MDS map produced from road distances data

this in more detail, by introducing the technique of Generalised Procrustes Analysis.

An extension of this example shows that the input data are not limited to actual physical distances. Table 1.2 shows the length of time taken to drive between the places used in the previous example. Here, the basic premise is that the closer two places are, the shorter the time taken to drive between them. Figure 1.2 shows the MDS plot obtained from these data.

The first thing to notice about the map in Figure 1.2 is that it is reflected about the horizontal axis, when compared with the map in Figure 1.1. There are also differences between the two maps with regards to the positions of the towns and cities, especially with regards to the location of Norwich. Physically, most of the places are joined by motorways, except for Norwich. Traffic on average tends to move faster on motorways, and so analysing time to drive has had the effect of moving Norwich further away

Town	Birm	Bris	Car	Leeds	Liv	Lond	Manc	Newc	Nor	Oxf
Birmingham	0	1.650	3.333	2.183	1.783	2.533	1.783	3.900	3.533	1.400
Bristol	1.650	0	4.600	3.583	3.050	2.367	3.050	5.300	4.450	1.483
Carlisle	3.333	4.600	0	2.367	2.083	5.467	2.117	1.367	6.167	4.417
Leeds	2.183	3.583	2.367	0	1.267	3.683	0.950	1.933	3.917	3.083
Liverpool	1.783	3.050	2.083	1.267	0	3.950	0.733	3.117	4.817	2.917
London	2.533	2.367	5.467	3.683	3.950	0	3.917	5.417	2.933	1.533
Manchester	1.783	3.050	2.117	0.950	0.733	3.917	0	2.783	4.350	2.883
Newcastle	3.900	5.300	1.367	1.933	3.117	5.417	2.783	0	5.383	4.817
Norwich	3.533	4.450	6.167	3.917	4.817	2.933	4.350	5.383	0	3.417
Oxford	1.400	1.483	4.417	3.083	2.917	1.533	2.883	4.817	3.417	0

Table 1.2: Time taken (hours) to drive between selected places in England
(www.theaa.com (2008))

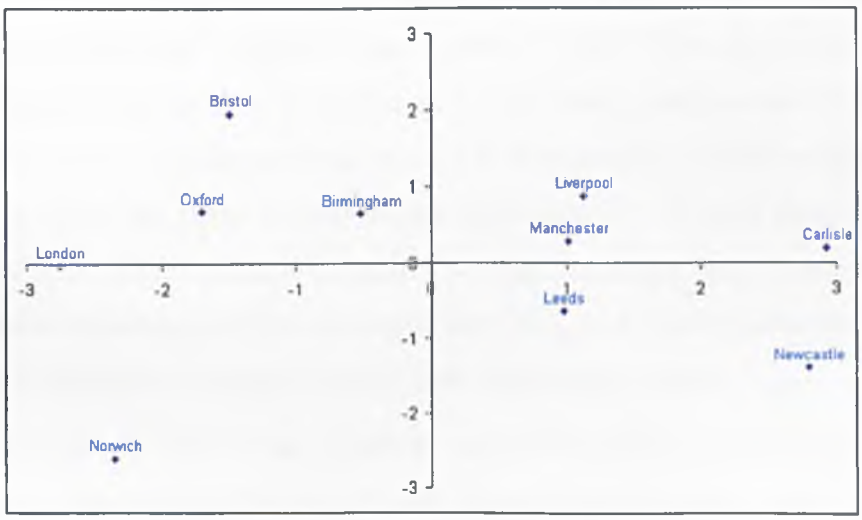

Figure 1.2: Metric MDS map produced from time taken to drive data

from the other places, especially the Northern cities.

1.3 The Theory of MDS

Although the analysis of the inter-city distances is a somewhat artificial example, it demonstrates the core idea underlying MDS: based on a dissimilarity measure (in these examples, physical distance or time taken) among a set of objects, MDS constructs a visualisation in which these objects appear as points on a map, and the closer two points are on the map, the nearer the objects are in multivariate space.

Mathematically, given O, a set of n objects, for each pair of objects $(i, j) \in O\text{x}O$ a dissimilarity measure δ_{ij} is defined - this being a non-negative number indicating how different or distant objects i and j are in some sense. If objects i and j are less alike than objects i' and j', then $\delta_{ij} > \delta_{i'j'}$. The MDS representation thus produced is a geometric configuration of points $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ in a p-dimensional space (often referred to simply as p-space) with $\mathbf{x}_i$ being the point corresponding to object $i \in O$, such that the set $\{d_{ij}\}$ of inter-point distances (usually Euclidean) matches 'as well as possible' the observed between-object dissimilarities $\{\delta_{ij}\}$, and thus reflects the inter-object information contained in the input dissimilarity data.

The phrase 'as well as possible' can be interpreted in many different ways, and this, along with the different ways of measuring d_{ij} and δ_{ij}, leads to the different types of multidimensional scaling.

The dimensionality p of the configuration is fixed before the configuration is produced (ideally with $p \leq 3$ to aid visualisation). In order to determine the most appropriate dimensionality in which to display the MDS

solution for a given set of dissimilarities, configurations can be produced for various values of p, and the optimum p can be chosen.

1.4 The input data - Proximities

The term proximities is often used with reference to the input data for multidimensional scaling. The definition of proximity is nearness in space, time, or some other way. Proximities refer to both dissimilarity and the opposite concept, similarity, with the obvious interpretation of measuring how dissimilar or similar objects are to each other.

Let the objects under consideration comprise a set O. The similarity/dissimilarity measure between two objects is then a real function defined on OxO, giving rise to similarity s_{ij}, or dissimilarity δ_{ij}, between the i^{th} and j^{th} objects. Usually $\delta_{ij} \geq 0$, $s_{ij} \geq 0$, and the dissimilarity of an object with itself is taken to be zero, i.e. $\delta_{ii} = 0$. Similarities are usually scaled so that the maximum similarity is unity, with $s_{ii} = 1$. A transformation such as $\delta_{ij} = (s_{ii} - 2s_{ij} + s_{jj})^{\frac{1}{2}}$ or $\delta_{ij} = s - s_{ij}$ (for some constant s) can be used to convert from similarities to dissimilarities. Without loss of generality, dissimilarities will be dealt with in the sequel.

There are many different possible dissimilarity measures that can be used for a set of objects (Cormack (1971), Snijders et al. (1990), Gower (1985), DeJordy et al. (2007)), which can broadly be classified into two groups, depending on whether the measure is computed or arrived at empirically. An example of a computed measure is where there is a set of t variables recorded on n objects (for example, the results of a sensory test), data which can thus be represented vectorially in terms of these measurements. In this case, a dissimilarity value can be based on some combination

of the vector components, for example *Euclidean distance* for differences between the objects:

$$\delta_{ij} = \left(\sum_{k=1}^{t} (x_{ik} - x_{jk})^2 \right)^{\frac{1}{2}}$$

or correlation-based, for differences between attributes:

$$\delta_{ij} = 1 - \frac{\sum_{k=1}^{t} (x_{ik} - \overline{x}_i)(x_{jk} - \overline{x}_j)}{\left(\sum_{k=1}^{t} (x_{ik} - \overline{x}_i)^2 \sum_{k=1}^{t} (x_{jk} - \overline{x}_j)^2 \right)^{\frac{1}{2}}}.$$

The earlier map distances example demonstrates an empirical production of dissimilarities. Another example is a card sort. Here, concepts and ideas about a product are written onto individual pieces of card, and sensory panellists are asked to cluster the concepts together, determined by how similar they think the concepts are. The number of clusters is left to the individual panellist to decide. The input data for MDS then takes the form of the number of times pairs of cards are placed into the same cluster - the higher this measure of co-occurence, the more similar the concept (Blake (2008)).

Choice of proximity measure depends upon the problem at hand, and is often not an easy task. Cox and Cox (2000) contains a review of various proximity measures, together with their associated problems.

There are four demands to be made of a measure of dissimilarity:

1. $\delta_{ii} = 0 \qquad \forall i$

   This captures the notion that an object is not dissimilar to itself in any way.

2. $\delta_{ij} \geq 0 \qquad \forall i \neq j$

   This means that two non-identical objects have a non-negative dissim-

ilarity between them, in the same way that two non-coincident points have a positive distance between them.

3. $\delta_{ij} = \delta_{ji} \qquad \forall i, j$

   This represents the fact that usually the order in which the objects under consideration are presented is irrelevant for the dissimilarity calculation. This symmetry of dissimilarities is not always evident, and theory exists to deal with asymmetric dissimilarity sets (Bove (2005)).

4. $\delta_{ij} \leq \delta_{ik} + \delta_{jk} \qquad \forall i, j, k$

   Here, the concept is that if two objects i and j are not too dissimilar from a third object k, then the dissimilarity δ_{ij} should itself not be large.

When a dissimilarity measure satisfies all four of these conditions, it is called a metric. Condition 4 is not always insisted upon.

1.5 MDS models

As mentioned in Section 1.3, the aim of MDS is to match the inter-point distances d_{ij} to the between-object dissimilarities δ_{ij} as closely as possible. There are many different ways for this matching to take place, each giving rise to a different type of MDS model. The most important division is into the classes of metric and non-metric methods. The main difference between the two is that in metric methods, the actual dissimilarity values themselves are important, whereas in non-metric methods, this is relaxed so that it is the rank order of the dissimilarities which is important.

In addition, it is worth highlighting here the problem that was encountered in Figure 1.1. Using most methods of MDS, the configuration of

points in p-space obtained from any given set of dissimilarities is by no means unique. In particular, any configuration can be translated, rotated and reflected, and still be a valid MDS representation of the information in the data. Furthermore, if a non-metric method is used, then the MDS solution is invariant with respect to uniform dilation/contraction also.

Borg and Groenen (2005) define an MDS model as a transformation f of dissimilarities δ_{ij} into distances d_{ij} in an MDS configuration in p-space, i.e.

$$d_{ij} = f(\delta_{ij}).$$

However, usually because of noise in empirical dissimilarities, and the use of iterative methods for minimising functions of several variables when actually finding such distances, it is necessary to relax the model to

$$d_{ij} \approx f(\delta_{ij}).$$

Borg and Groenen (2005) list several types of function giving rise to metric MDS models. For example, absolute MDS is where f is the identify function:

$$d_{ij} = \delta_{ij}.$$

Alternatively there is ratio MDS

$$d_{ij} = b\delta_{ij} \qquad (b > 0)$$

and interval MDS

$$d_{ij} = a + b\delta_{ij} \qquad (a, b > 0).$$

The function f need not be linear, for example

$$d_{ij} = a + b\log(\delta_{ij}) \qquad (a, b > 0)$$

and

$$d_{ij} = a + b\exp(\delta_{ij}) \qquad (a, b > 0).$$

Non-metric models require only a monotonic function such that

$$d_{ij} < d_{kl} \Leftrightarrow f(\delta_{ij}) \leq f(\delta_{kl}).$$

Non-metric MDS methods are far more widely used in practice than metric methods, as often (especially for empirically collected dissimilarity data) all that is really of worth are the rank orders.

For more details of two of the main MDS models, see Appendix A and Appendix B.

1.6 Procrustes Analysis

The situation may arise when two MDS configurations are available for a given set of objects. These could have arisen, for instance, by performing MDS via two different techniques, such as classical scaling and non-metric scaling. Alternatively, an experiment could have been repeated on the same objects at different times, or there could simply be two sets of measurements such as the distance and time measurements in Section 1.2. However they arise, there exists a one-to-one correspondence between the points in the two configurations.

Procrustes analysis (least-squares orthogonal mapping) was initially developed for use in factor analysis (Mosier (1939) and Green (1952)), and

its use in multidimensional scaling has been developed in papers such as Schönemann and Carroll (1970) and Gower (1971). Procrustes analysis is commonly used in statistical shape analysis, investigating the distributions of a set of shapes. Reviews of Procrustes analysis can be found in Kendall (1989) and Gower and Dijksterhuis (2004).

Procrustes analysis finds the best match between two configurations. Simply put, the method is based on matching corresponding points (or landmarks) from each of the two data sets. The objective is minimise the sum of the squared deviations (termed the error, and denoted as the m^2 term) between landmarks through translating, rotating and dilating one configuration to match the other configuration (i.e. the target). The deviations between landmarks are called vector residuals - a small vector residual indicates a close agreement between the corresponding landmarks.

Mathematically, consider n points in two-dimensions, say $[(x_{11}, x_{21})$, (x_{12}, x_{22}),...,$(x_{1n}, x_{2n})]$. The mean of these points is $(\overline{x}_{1.}, \overline{x}_{2.})$, where $\overline{x}_{i.} = n^{-1}\sum_{j=1}^{n} x_{ij}$; $(i = 1, 2)$. Now the configuration is translated so that the mean is the origin $(x_{1i}, x_{2i}) \rightarrow (x_{1i} - \overline{x}_{1.}, x_{2i} - \overline{x}_{2.})$. Likewise, the scale can now be removed by finding the size of the configuration

$$s = \sqrt{\sum_{i=1}^{n} \left[(x_{1i} - \overline{x}_{1.})^2 + (x_{2i} - \overline{x}_{2.})^2\right]}$$

and dividing the coordinates by the scale, s. There are also other methods for removing the scale that can be used.

To remove the rotational component, consider two configurations, that have had their scale and translation removed, with points (x_{1i}, x_{2i}) and (y_{1i}, y_{2i}). Let the $\mathbf{X}$ configuration be fixed, and the $\mathbf{Y}$ configuration be rotated around the origin, so that the sum of the squared distances between

the points is minimised. Let (y_{1i}, y_{2i}) be mapped to (u_{1i}, u_{2i}) through a rotation by angle θ. Thus $(u_{1i}, u_{2i}) = (y_{1i}\cos\theta - y_{2i}\sin\theta, y_{1i}\sin\theta + y_{2i}\cos\theta)$. The Procrustes statistic, or distance, is

$$d = \sqrt{\sum_{i=1}^{n} [(u_{1i} - x_{1i})^2 + (u_{2i} - x_{2i})^2]},$$

which is minimised by using a least squares technique to find the angle θ that gives a minimum distance. The distance d provides a metric to measure how close the two configurations match each other.

Generalized Procrustes analysis (GPA) is a procedure applying Procrustes analysis to align more than two configurations. It can also be used to find the 'average' configuration of a set of configurations.

The following example shows the technique of generalized Procrustes analysis, and is based on the example in Section 1.2. Figure 1.3 shows non-metric MDS maps of the Distance and Time to Travel data from Tables 1.1 and 1.2. The aim is to produce an average configuration of the two configurations.

Procrustes analysis results in the plot in Figure 1.4. The points in blue represents the MDS configuration on the distance data, the points in green are the MDS configuration on the time data, whilst the points in red are the consensus configuration from the Procrustes analysis. As can be seen, the point for Norwich has altered considerably, which confirms the earlier detail about the lack of motorways impacting on the time taken to drive. However, the points for Newcastle, Carlisle and Bristol have also altered. This could be for any number of reasons.

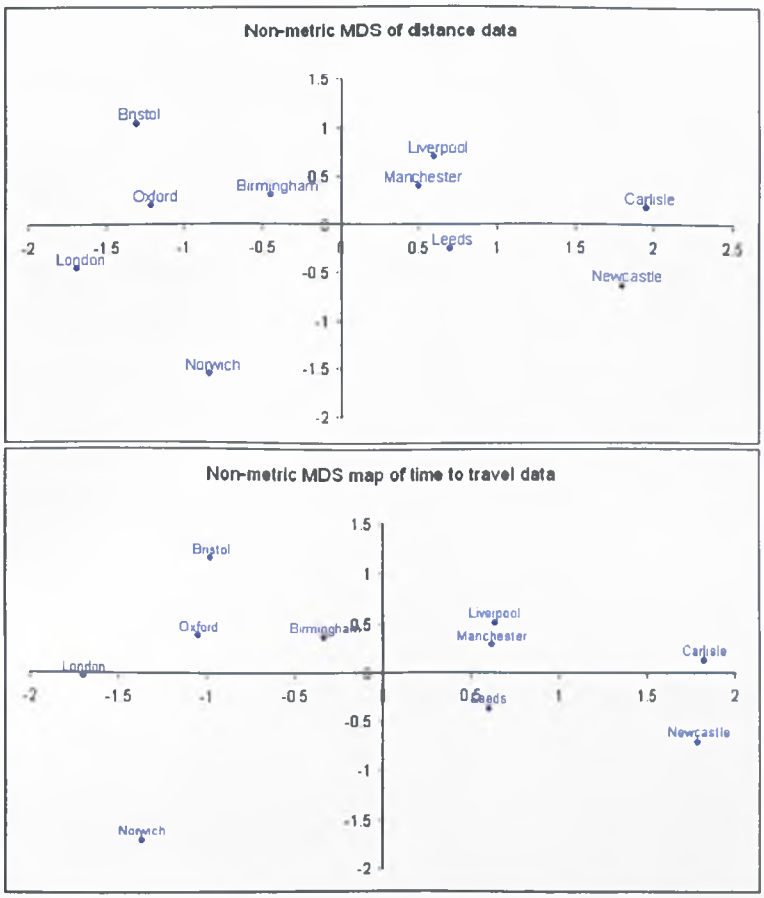

Figure 1.3: Non-metric MDS map based on distances, and non-metric MDS from time to travel data

1.7 Extensions to Multidimensional Scaling

1.7.1 Adding information to MDS plots

As mentioned previously, it can be difficult to interpret multidimensional scaling plots. The map example of Section 1.2 could be interpreted using a basic knowledge of geography. However, in practice it can be much more

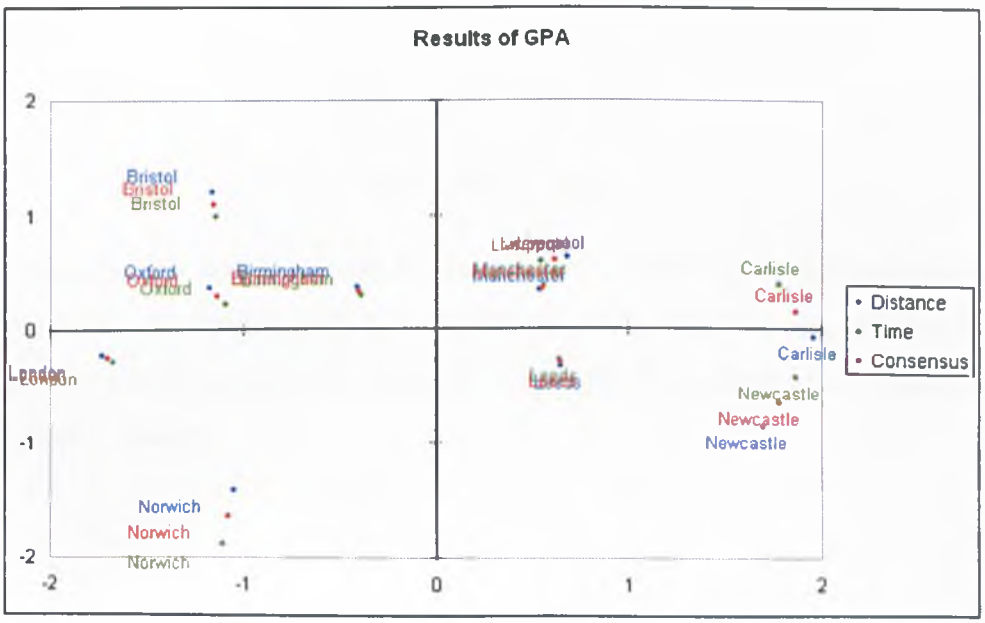

Figure 1.4: Results of GPA on distance and time MDS maps

difficult to interpret 'real-life' plots.

A simple method for finding meaningful directions or axes within the configuration is to use multiple linear regression (Kruskal and Wish (1978)). An axis is found for a variable related to the objects. This variable, which can be called y say, is taken as the dependent variable, with the coordinates of the points in the configuration being the independent variables.

The regression model is then

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\mathbf{y}$ is the vector of observations $\{y_i\}$ $(i = 1, \ldots, n)$, $\mathbf{X}$ is the $nx(p+1)$ matrix consisting of a column of ones followed by the coordinates of the points in the configuration, $\boldsymbol{\beta}$ is the parameter vector, and $\boldsymbol{\epsilon}$ is the 'error'

vector.

The least squares estimate of β is given by

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

As long as the regression has a reasonable fit, tested either by an analysis of variance, or by the multiple correlation coefficient, then an axis for the variable can be defined through the origin of the configuration using the direction cosines

$$\frac{\hat{\beta}_i}{\sqrt{\sum \hat{\beta}_i^2}} \qquad (i = 1, \ldots, p).$$

1.7.2 Biplots

Biplots are statistical graphs that represent variables and objects in the same plot. Biplots can be seen as the multivariate analogue of scatter-plots: they give a graphical representation of a multivariate sample and they superimpose on the display a representation of the variables on which the sample is measured. Alternatively they can be thought of as a way of adding information about the attributes used to multidimensional scaling plots, in an attempt to aid interpretation.

In a biplot, objects are displayed as points (as with MDS), while the attributes are displayed either as vectors, linear axes, or non-linear trajectories. In the case of categorical variables, category level points may be used to represent the different levels of the variable, giving rise to a generalised biplot which displays information on both continuous and categorical variables. The biplot was introduced by Gabriel (1971) and developed further by others, but especially by Gower and Hand (1996). Recent examples and developments can be found in the work by Park et al. (2008), Pittelkow and

Wilson (2007) and Friendly (2007), whilst Udina (2005) develops an interactive biplot graph.

For the continuous variables only case, the mathematics involve a singular value decompostion of the data. Let $\mathbf{X}$ be an (nxk) matrix of data, representing n objects measured on k attributes, where n k. The singular valued decomposition of $\mathbf{X}$ is given by

$$\underset{nxk}{\mathbf{X}} = \underset{nxk}{\mathbf{U}}\ \underset{kxk}{\mathbf{L}}\ \underset{kxk}{\mathbf{V}'}$$

where the diagonal matrix $\mathbf{L}$ contains the singular values orderd by magnitude down the diagonal, $\mathbf{U}$ contains the corresponding left singular vectors, and $\mathbf{V}$ contains the corresponding right singular vectors.

For a two-dimensional biplot, the SVD results are used to form the 2x2 matrix $\underline{\mathbf{L}}$, which contains the two elements of $\mathbf{L}$ with the highest singular values. The nx2 matrix $\underline{\mathbf{U}}$ and the kx2 matrix $\underline{\mathbf{V}}$ are formed by choosing those columns from $\mathbf{U}$ and $\mathbf{V}$ which correspond to these highest singular values. The coordinates for the observations are given by

$$\underset{nx2}{\mathbf{G}} = \underline{\mathbf{U}}\underline{\mathbf{L}}^{c}$$

and

$$\underset{nx2}{\mathbf{H}'} = \underline{\mathbf{L}}^{1-c}\underline{\mathbf{V}}'$$

where the value of c defines the type of biplot:

- GH-Biplot: $c = 0$. This preserves correlations between the variables and allows projection of the objects onto the vectors.
- JK-Biplot: $c = 1$. This preserves distances between the objects (cf. MDS) and allows projection of the objects onto the vectors. This type of biplot is equivilent to a PCA on the data.

- SQ-Biplot: $c = 0.5$. This preserves distances and correlations, but the projection property is lost.

The objects are represented by the points with coordinates given by the rows of $\mathbf{G}$, whilst the coordinates for the variables are given by the rows of $\mathbf{H}$.

Example

A sensory test was carried out on the application properties of eight aerosol anti-perspirant deodorants. The trained panel were asked to score several attributes on a 100-point scale. A summary of the results is given in Table 1.3. An Analysis of Variance was carried out on the data, along with Tukey HSD multiple comparison tests to highlight the nature of any significant differences. In Table 1.3, products that are not significantly different on an attribute at the 95% level of confidence have been given the same letter.

	A		B		C		D	
Loudness of spray	75.03	a	74.09	a	61.81	c	69.00	b
Evenness of spray	94.31	a	64.19	a	64.31	a	65.28	a
Ease of use of spray	73.81	a	74.06	a	73.78	a	74.09	a
Strength of spray	74.81	a	74.87	a	63.84	c	70.81	ab
Directability of spray	74.34	ab	75.03	a	73.47	abc	73.16	abc
Coldness of spray	70.41	b	69.22	bc	61.88	de	68.72	bc
Visibility of product on skin	39.16	b	23.31	c	18.56	c	22.97	c
Wetness on skin	19.88	cd	23.91	bc	20.63	cd	31.66	a
Stickiness on skin	13.16	cd	15.09	bc	13.34	cd	18.28	a
Greasiness on skin	18.44	bc	20.81	b	20.59	b	26.19	a

	E		F		G		H		p-value
Loudness of spray	67.28	b	67.62	b	67.03	b	65.84	bc	< 0.0001
Evenness of spray	64.47	a	64.75	a	65.41	a	64.28	a	0.5142
Ease of use of spray	74.09	a	73.90	a	73.68	a	72.84	a	0.1112
Strength of spray	67.03	bc	68.16	bc	70.97	ab	65.31	c	<0.0001
Directability of spray	71.44	bc	70.91	c	74.41	ab	73.31	abc	0.0004
Coldness of spray	64.88	cd	68.63	bc	74.97	a	58.59	e	<0.0001
Visibility of product on skin	21.53	cd	22.78	c	18.81	c	45.16	a	<0.0001
Wetness on skin	27.19	ab	28.66	ab	28.47	ab	16.13	d	<0.0001
Stickiness on skin	16.44	ab	19.00	a	13.81	bcd	11.22	d	<0.0001
Greasiness on skin	26.16	a	26.25	a	17.88	bc	15.28	c	<0.0001

Table 1.3: Mean scores from sensory study on deodorants, along with ANOVA results

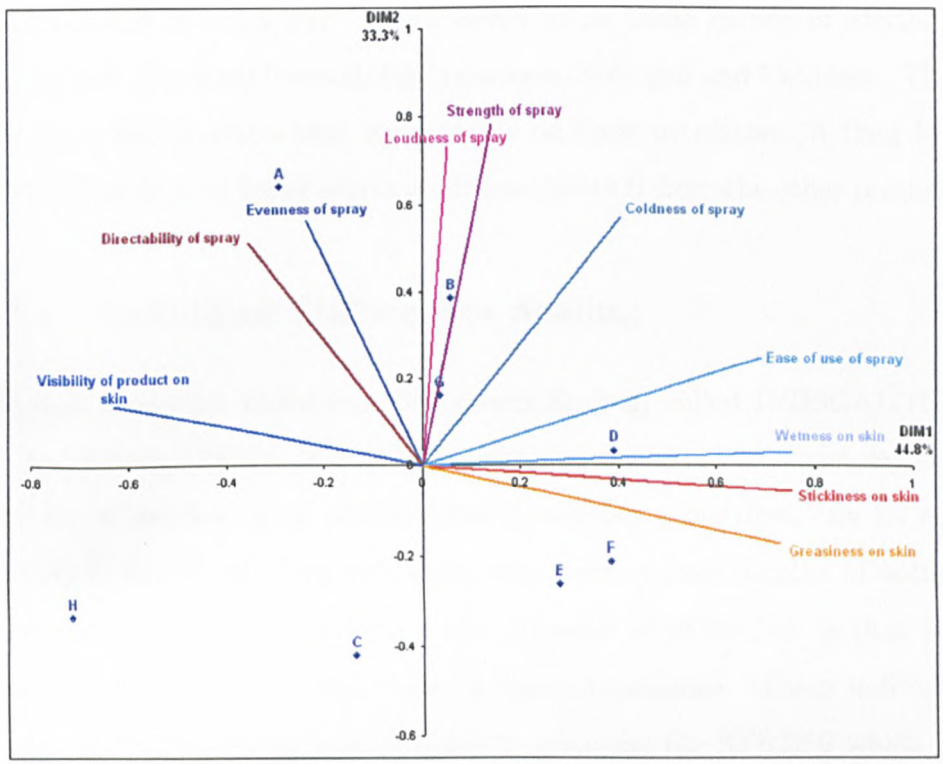

Figure 1.5: Biplot of sensory data

Figure 1.5 shows the GH biplot produced from these data.

The biplot is based on PCA, and so can be interpreted as follows. Dimension 1 is dominated by how the product feels (or looks) on the skin. Products towards the right hand side of the plot (D, E and F) are those which initially are wet on the skin (with the concepts of stickiness and greasiness being closely related to wetness), whereas towards the left is product H which is highly visible. In sensory testing of AP deodorants, visibility is defined as white marks, which tend to be dry and powdery. Thus dimension 1 can be thought of as dry to wet feel of application.

Dimension 2 differentiates the products on the physical application

characteristics of the spray. There seems to be three groups of attributes driving this: Eveness/Directability, Loudness/Strength and Coldness. There are only a few products that score highly on these attributes (A then B, G and D). The Ease of Spray attribute differentiates H from the other products.

1.7.3 Individual Differences Scaling

The first model for Individual Differences Scaling, called INDSCAL (Carroll and Chang (1970)), is an MDS technique dedicated to three-way data analysis. It analyses a set of individual dissimilarity matrices, one for each individual. INDSCAL iteratively defines an *a priori* fixed number of optimal dimensions mapping the objects. The strength of INDSCAL is that each assessor can weight differently each of these dimensions. These individual vectors of weights are defined in order to minimise the STRESS which is a least square criterion between the observed individual dissimilarities and the compromise object distances in the fitted space. Subsequent individual differences scaling approaches have been developed by, amongst others, Young and Hamer (1987) and Krzanowski (1988). They form a group of techniques known as weighted MDS (WMDS) (Schiffman et al. (1981)). WMDS analyses several data matrices at the same time - each matrix represents the results of a separate experimental condition, a separate individual, or group of individuals.

In WMDS, differences among individuals are reflected as differences in weights for a set of common underlying dimensions. In addition to a group stimulus space (or consensus spatial configuration), WMDS derives dimension weights for each individual that can range from 0 to 1 and reflect the relative importance of each dimension to the individual.

WMDS shares many features in common with classical MDS. WMDS can be metric or nonmetric. The degree of fit is evaluated in the same fashion, except that there are measures of fit for the group space as well as the individual spaces. However, there is one technical difference. As seen previously, the dimensions in MDS can be rotated. But this is not the case for WMDS. This means that the dimensions in WMDS can possibly be interpreted. Schiffman et al. (1981), however, point out that this non-rotatability is true strictly only when the data contain no error. In the presence of error, some amount of rotation is permissable.

Finally, it should be noted that WMDS is based on a particular view of individual differences, namely that individuals differ in the relative importance they assign to a set of common dimensions. This is the only point of difference among individuals, according to the WMDS model. Mathematical extensions of WMDS models (see Young and Hamer (1987)) include differences among individuals in rotation of the group space and in the number of dimensions of the personal spaces. These extensions of the basic WMDS model, however, have found relatively few applications to date.

1.8 Applications of MDS

The aim of this Section is to demonstrate the many different areas and applications of Multidimensional Scaling. It is by no means an exhaustive discussion, but instead aims to give an idea of the possibilities of the technique, and some of the recent developments in the methodology.

As mentioned in Section 1.1, Multidimensional Scaling has found uses in many different fields. Perhaps of most relevance to this thesis, is its use in the areas of Consumer and Sensory science. For example, a study by Lim

and Green (2007) looked at the ability of capsaican (often seen as a pure sensory irritant) to evoke, mask and desensitize bitter taste, suggesting that the burning sensation from capsaican might perceptually be related closely to the bitter taste. A sensory test was carried out looking at taste stimuli, and a MDS map from these data showed that capsaican was similar to quinine sulphate (a very bitter compound) despite being mediated by different sensory modalities. Also this 'bitter-burning' group was clearly separated from the other taste stimuli (sweet, sour and salty), which may be due to their common function as sensory signals of potentially harmful stimuli. Another sensory study (Yoshioka et al. (2007)) investigating texture perception used MDS to investigate the similarities between a probe and the human finger on a variety of textured surfaces. The results showed that the two methods were similar though not identical - roughness ratings being near identical, but hardness and stickiness differing. Meanwhile, an investigation into what attributes are the major determinants of aesthetic appeal of photographs used MDS to yield 3 dimensions that drive people's perception of photographs (Axelsson (2007)). With the aid of attribute scales combined with measures of the manifest content of the photographs, it was possible to identify these dimensions as Hedonic Tone-Familiarity, Absence of Colour, and Expressiveness-Dynamics. Finally, an adaptation to MDS known as Probabilistic MDS provides a mechanism for accounting for the variability inherent in sensory data by using distributions, instead of points, to represent sensory objects (MacKay and O'Mahony (2002)).

Giragame et al. (2006) used MDS to generate a perceptual structure map for colour tone stimuli responses from naive television viewers from two different cultures.. The map revealed that Japanese and Sinhala native speakers perceptually discriminate the six colour tone stimuli into six differ-

ence categories. Furthermore, the semantic structures used for each hue were different from one another. Keeping with television, Nabi (2007) used MDS to find that there were two underlying dimensions along which audiences think about reality TV, based on data from a sorting task of thirty-three reality-based programs. These two dimensions were romance and competitiveness.

Looking at mental object representations, Cooke et al. (2007) developed MDS to find that a single underlying perceptual map (with dimensions corresponding to shape and texture) could explain visual, haptic and bimodal similarity ratings on novel 3-dimensional objects.

MDS is very commonly used in the analysis of consumer data. Martins and Pliner (2006) carried out a study aimed at identifying what characteristics of food made individuals perceive them as disgusting. A non-metric MDS map indicated two dimensions - i) aversive textural propery of the food; and ii) reminders of livingness/animalness - accounted for most of the variability in the consumer scores. Meanwhile Cunningham (2006) used MDS to show that a 2 dimensional solution, that was identical across cultures, for describing how customers perceived and classified a set of Governmental services. Kagie et al. (2007) used MDS to develop a new graphical user interface for online shopping that uses a map of the product space to show consumers similar products, allowing consumers to improve their choice of the product range. This was demonstrated with an on-line shop for MP3 players. Consumer opinions about on-line travel agencies were investigated by Kim et al. (2007), where MDS was used to display similarities and patterns based on travellers' perceptions in terms of web features, user friendliness, security and cost. Continuing with consumer perception, MDS was used by Zheng et al. (2007) to see how perceived and actual perception are linked, and the

effect of product appearance on the perceived useability of car infotainment systems, whilst Petiot and Grognet (2006) carried out a similar study on cars, and linked this into the product design process.

There have been several extensions to MDS to handle different aspects of analysing consumer data. In Chen et al. (2008), a weighting is applied to the objects used to produce the MDS map in an attempt to match differences in the weighting that people give to subjective/perceptive judgements in the real world. DeSarbo et al. (2006) details a stochastic MDS that is calibrated from actual consumer consideration/choice sets to estimate and uncover competitive market structures that are asymmetric, whilst Faye et al. (2006) use MDS as an alternative to external preference mapping through two consumer test phases (a preference scaling and a perceptual free sorting followed by verbal description). This global approach allows preference to be explained by the consumer perceptual dimensions and these dimensions to be interpreted using the words that consumers use. In fact the application of MDS to consumer semantics is a very common procedure - Lin et al. (1996) and Verheyen et al. (2007) both use the words of consumers to build psychological spatial representations that are used to aid product designers.

One of the biggest areas in which MDS is applied is in genetics, and the study of DNA and genomes. Alfonso-Sanchez et al. (2008) uses MDS to map the genomic diversity of the Arrento people of Australia, and finds that they are closely related to people from the Indian subcontinent. Thai et al. (2007) does a similar thing, but with the common carp in Vietnam, whilst Kar et al. (2008) separates the Mulberry germplasm based on genetic diversity and protein content. Finally Zhou et al. (2007) uses MDS to disprove a theory that the Liqian people of Northern China were descended from the Romans - they are in fact more closely related to a Chinese subgroup called

the Han.

The genetics area also extends to Bioscience. Oh and Raftery (2007) apply this to gene expression data on genes believed to be informative about the distinction between two forms of leukemia, and were able to cluster the two sets of genes. Meanwhile Napolitano et al. (2008) used MDS to produce a 2-dimensional visualization of human cell cycle gene expression data, and use this to identify genes periodically expressed in a human cancer cell line. Amaratunga et al. (2008) used an extension of MDS for visualising DNA microarray data. MDS was also used to compare the different methods for reducing the alphabet of 20 amino acids involved in protein structures, thereby aiding the investigation of amino acid interactions (Luthra et al. (2007)).

Staying with biological data, many researchers have used MDS to map ecological data. Wright et al. (2005) used non-metric MDS to show differences in foliose algal community composition on temperate marine reefs due to grazing by sea-urchins, whilst Acevedo and Restrepo (2008) used the same method to show that land use followed by climate could explain most of the variation observed among the composition and abundance of birds in Puerto Rica. Moreover, endemic and exotic species were widely distributed throughout the island, but the proportion of endemic species was higher in closed forests, whilst exotic species were more abundant in open habitats. Also working in this area, Della Bella et al. (2008) used MDS on the taxonomic composition of aquatic plants between temporary and permanent ponds in central Italy, and Butler and deMaynadier (2008) looked at the diversity and composition of damselfly assemblages and related these to anthropogenic degradation in excessively developed waterbodies, using MDS. Finally, non-metric MDS is mentioned in a review of appropriate multivariate methods

for Environmental Survey data and Biotic (Species) Survey data by Kenkel (2006).

MDS is used in a wide variety of other applications too. Here, a small selection of examples is discussed. Mugavin (2008) has produced an introduction of MDS as a technique for studying medical patients' perception of cancer pain, breathlessness in individuals with chronic obstructive pulmonary disease, and the assessment of vulnerable populations where social desirability is an issue. Staying with medicine, rules are tools used in the profession to diagnose disease. MDS has been applied to rule induction methods to show similarities between the rules generated from large datasets. This method gave experimental results useful for domain experts (Tsumoto and Hirano (2007)).

In crime studies, MDS was used by Dixon et al. (2008) to construct a classification system of men who are incarcerated for the murder of their female partner. MDS identified three sub-groups - low criminality/low pyschopathology, moderate-high criminality/high pyschopathology, and high criminality/low-psychopathology, which can be used to determine appropriate treatment.

A free-sorting method coupled with MDS suggested that thinking about college student types in the US should include academic involvement and social involvement dimensions (Ashmore et al. (2007)). The MDS showed that positive and negative social, positive academic, and oppositional clusters of types were seen in studies of high school students, and that the cognitive structure underlying perception of college student types was converging across major demographic categories.

MDS is useful for determining the position of a mobile station in a wireless communication system (Chen et al. (2008)). Distances are measured

between base stations (with known positions) and the mobile stations, using time-of-arrival measurements, and MDS used to locate the mobile stations. A dynamic approach is used to improve performance by combining location information from measurements made at several sampling time points.

MDS is widely used in shape analysis. Both Liu et al. (2007) and Lespinats et al. (2007) use MDS as the basis of a nonlinear dimension reduction method and demonstrate the principles on facial expression and recognition techniques.

In Matheus et al. (2006), industrial process monitoring involving the collection of real-time multivariate data (often $\sim 10^{18}$ bytes of data produced every year for one process) is analysed using MDS with an iterative capability. Projected Orientation Mapping allows newly obtained data points to be added to the existing maps in real-time, so the operational regions of the process under specific conditions can be easily classified. This new methodology was applied to the oil industry. Interactive graphs were also used on social networking data in Hosobe (2007)

Profile Analysis via Multidimensional Scaling (PAMS) provides an exploratory technique for visualising profile data. In the paper by Ding (2007), PAMS provided an exploratory technique for identifying major growth profiles by extending the model for longitudinal data. The MDS profile model was solved for the growth parameters such that each MDS dimension corresponded to a major growth profile. This led to an identification of the growth trends, allowing for a study of the individual differences with respect to those growth trends. Meanwhile, Kim et al. (2007) developed a Confirmatory Factor Analysis parameterisation of the PAMS model to demonstrate validation of a profile pattern hypothesis derived from MDS. This was applied to the General Occupation Theme survey, and looked at drivers of interest behind

people's opinions of their careers.

In Astronomy, Lilensten et al. (2007) used MDS to select the best set of lines that could be used to reconstruct the solar Ultra-Violet spectrum. This allowed improved monitoring of solar irridiance, which is a crucial issue in space weather forecasting.

Finally, MDS has been used to estimate the positions and postures of a demonstrater which is to be mimiced by a humanoid robot (Lee and Nakamura (2007)).

Chapter 2

Adding Curved Axes to MDS Maps

2.1 Introduction

As seen in Section 1.2, the interpretation of Multidimensional Scaling maps can be quite difficult. The map featured in Figure 1.2, showing a plot based on time to travel, is a case in point. It is known from basic UK geography that Newcastle is further North than London, and that Bristol is further West than Norwich. Thus it can easily be determined that the right hand side of the Figure represents North, and the top of the plot is West. However, this is a relatively trivial example. In practice, the interpretation is much more difficult, as the different dimensions could represent several attributes concurrently. Alternatively, the dimensions may not represent anything - they are just an artifact of the construction of the map after all. Most MDS maps are invariant to rotation and translation making the coordinate axes meaningless. There are some exceptions though, for example Individual

Differences Scaling does have unique axes (Carroll and Chang (1970) and Young and Hamer (1987)).

As described in Section 1.7, there are several mechanisms for attempting to interpret MDS maps. The most simple method is to use multiple linear regression to fit an axis for a variable, which is taken as the dependent variable(Kruskal and Wish (1978)). The independent variables are the coordinates of the points in the final configuration.

Alternatively, biplots attempt to show not only the configuration of points representing the objects, but also axes within the plots that represent the variables upon which the original measures were made (Gower and Hand (1996)). In the simplest case, the axes are linear, but with generalisation the axes can be non-linear. However, to highlight one of the limitations with biplots - generally the variable that forms the axis must have been used to define the map. In the driving time example, there are no initial variables with which to construct a biplot - the input data are simply the times taken to travel between the cities.

What is needed is a method whereby additional information about the objects could be used to define axes that add meaning to the configuration. In the driving time example, once the map has been constructed, an axis based on the latitude of each city could be overlaid. This would indicate the North/South direction on the map - the greater the latitude, the further north a place is. A similar axis can be overlaid based on the longitude value for the East/West direction - see Figure 2.1. An axis has been fitted by projecting each object onto it, to give a projection score on the axis for that object. The parameters of the equation defining the axis are then adjusted in order to minimise the differences between the projection score and the actual observed score for each object.

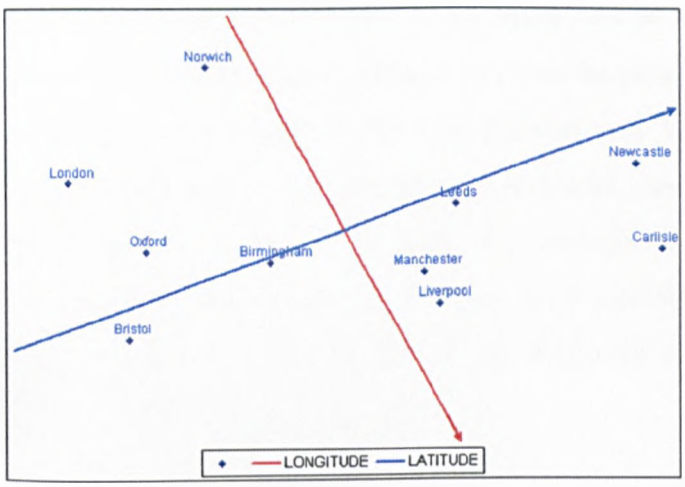

Figure 2.1: Example of MDS map for driving time data, with axes overlaid for longitude and latitude values. Axes point towards North and West

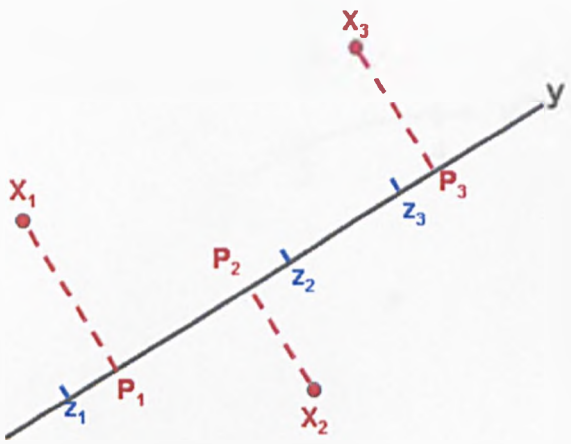

Figure 2.2: Schematic showing the principle for fitting an axis y to an MDS map of three objects X_i; P_i are the projected scores of the objects onto y, whilst z_i are the observed values of object X_i on the attribute that is represented by y.

Figure 2.2 shows a schematic based on an MDS plot of three objects and a linear axis. This shows how a straight line can be fitted (analogous with the multiple linear regression method of Kruskal and Wish (1978)). However, it might sometimes be more useful to deal with cases where the optimum value is not at the end of the scale. An example of this are the so-called JAR scales in sensory science, where the scale runs from too little of a quantity through the optimum of Just About Right, to too much of a quantity (Gacula et al. (2007))

2.2 Adding Axes to Multidimensional Scaling plots

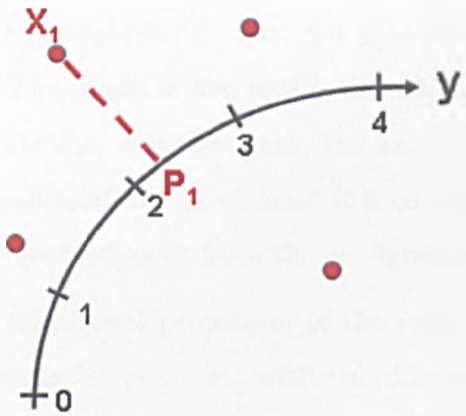

Figure 2.3: Schematic showing projection of MDS points onto axis

Figure 2.3 shows a curved axis for an attribute y, fitted to an MDS plot. The axis has to have an origin defined, and a scale of measurement. The value read on the axis for a point X_i is the distance along the axis from the origin to the point where X_i projects orthogonally onto the axis, say l_i.

Let z_i be the actual recorded value for the attribute associated with point X_i. The curved axis then is fitted by minimising

$$L = \sum_{i=1}^{n} (l_i - z_i)^2 . \tag{2.1}$$

So far, the defining form of the axis has not been given. There is an unlimited choice of functions that could be used for its definition. In the next Subsection, polynomials of the Cartesian coordinates are used to demonstrate the procedure.

2.2.1 Fitting a Curved Axis

This Subsection details an axis defined by a quadratic polynomial of the Cartesian coordinates. Let the axis be parameterised by, for example, $x_1(t) = t, \quad x_2(t) = b_0 + b_1 t + b_2 t^2; \quad -\infty \leq t \leq \infty$ with the origin of the axis being when $t = 0$. The origin of the axis is not necessarily the same as the origin of the configuration. For this work, the axis has been forced through the origin of the configuration, as without this constraint, it is possible for the fitted axis to be located away from the configuration. Thus, here $b_0 = 0$.

Now, let the orthogonal projection of the point X_i (with coordinates (x_{1i}, x_{2i})) onto the axis be point P_i, with coordinates (x_{1i}^p, x_{2i}^p). Then the value read from the curved axis for point X_i is the length of the curve from the origin to point P_i, say l_i, as in Figure 2.4.

For the axis $(x_1(t), x_2(t))$, the equation of the normal to the tangent at any one particular point is

$$x_2 = -\frac{1}{\beta} x_1 + c \tag{2.2}$$

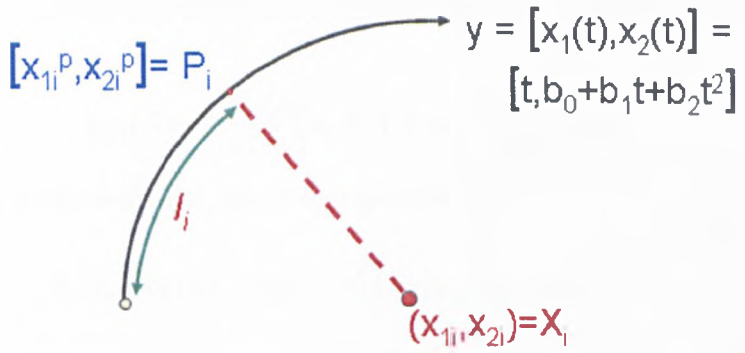

Figure 2.4: Schematic showing the finding of the projected value l_i

where β is the slope of the tangent.

As the normal line must past through X_i, which is (x_{1i}, x_{2i}) then

$$x_{2i} = -\frac{1}{\beta}x_{1i} + c$$

or

$$c = x_{2i} + \frac{1}{\beta}x_{1i}. \tag{2.3}$$

Now define t_i as the value of t at point P_i. Then the slope of the tangent is given by

$$\beta = \frac{x_2'(t_i)}{x_1'(t_i)}.$$

This, along with (2.3), can be substituted into (2.2), giving

$$\begin{aligned} x_2 &= -\frac{1}{\beta}x_1 + c \\ &= -\frac{1}{\beta}x_1 + x_{2i} + \frac{1}{\beta}x_{1i} \\ &= -\frac{x_1'(t_i)}{x_2'(t_i)}x_1 + x_{2i} + \frac{x_1'(t_i)}{x_2'(t_i)}x_{1i}. \end{aligned} \tag{2.4}$$

As the normal line also passes through P_i then substitution into (2.4) gives

$$x_2(t_i) = -\frac{x_1'(t_i)}{x_2'(t_i)} x_1(t_i) + x_{2i} + \frac{x_1'(t_i)}{x_2'(t_i)} x_{1i}$$

which, after rearrangement, gives an equation for t_i:

$$x_2'(t_i)\left[x_2(t_i) - x_{2i}\right] + x_1'(t_i)\left[x_1(t_i) - x_{1i}\right] = 0. \tag{2.5}$$

Now, as the axis $[x_1(t), x_2(t)]$ was previously defined in the example as $[t, b_0 + b_1 t + b_2 t^2]$, then $x_1'(t) = 1$ and $x_2'(t) = b_1 + 2b_2 t$. So, substitution of these values into (2.5) gives

$$(b_1 + 2b_2 t_i)\left[\left(b_0 + b_1 t_i + b_2 t_i^2\right) - x_{2i}\right] + (1)\left[t_i - x_{1i}\right] = 0,$$

$$t_i^3\left[2b_2^2\right] + t_i^2\left[3b_1 b_2\right] + t_i\left[b_1^2 + 2b_0 b_2 - 2b_2 x_{2i} + 1\right] + \left[b_1 b_0 - b_1 x_{2i} - x_{1i}\right] = 0,$$

which is a cubic equation. The solution to this cubic equation can have one real root, two real roots, or three real roots. These correspond to the three positions of points, A, B and C respectively in Figure 2.5.

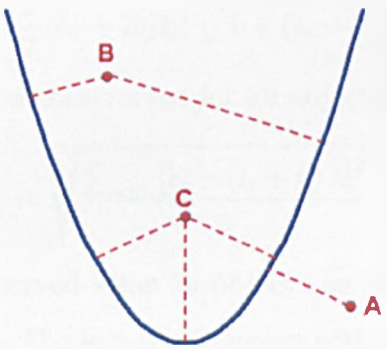

Figure 2.5: Possible projections to a quadratic axis

When there are two or more possible values for t_i, the t_i is chosen as the one which gives the shortest Euclidean distance between X_i and P_i, which in Figure 2.5, the projection to the left hand side would be selected for object B.

Once t_i has been found, the distance from the origin of the axis to P_i can be found using the standard equation for the length of an arc. Thus the value l_i is given by

$$l_i = \int_0^{t_i} \sqrt{x_1'(t)^2 + x_2'(t)^2}\mathrm{dt} + k, \tag{2.6}$$

where k is a constant which allows for the origin of the scale to be located anywhere along the axis.

Thus, for the example,

$$l_i = \int_0^{t_i} \left[1 + (b_1 + 2b_2 t)^2\right]^{\frac{1}{2}} \mathrm{dt} + k$$

and so

$$l_i = \frac{1}{2}\ln\left(\sqrt{1 + (b_1 + 2b_2 t_i)^2} + b_1 + 2b_2 t_i\right) + \frac{1}{2}(b_1 + 2b_2 t_i)\sqrt{1 + (b_1 + 2b_2 t_i)^2} + k$$

A loss function, L, is then constructed for all objects $i = 1, \ldots, n$ giving

$$L = \sqrt{\frac{\sum_{i=1}^{n} \left[z_i - (l_i + k)\,\lambda\right]^2}{n}}, \tag{2.7}$$

where z_i is the actual observed value for object i on the attribute to be fitted and λ is a scaling factor. The loss is minimised with respect to λ, k and the parameters defining the axis (b_0, b_1, and b_2 here). The loss can be thought of as a measure of the average difference between calculated and observed values on the axis.

It is also possible to remove scale from the attributes by dividing the loss function by the attribute standard deviation. This allows for the comparison of different attributes that have been measured on differing scales.

2.3 Examples and Applications

In this section, the methodology is demonstrated on simulated and real data sets.

2.3.1 Adding longitude and latitude axes to the city distance map

Table 2.1 shows the longitude and latitude measures for the cities featured in the example in Section 1.2.

An axis to indicate North would be expected to run from left to right on the map, whilst an axis for West would run from top to bottom.

Two axes were thus fitted to the MDS plot to represent the two variables (latitude and longitude). Initially, the axes were chosen to be linear, $(t, b_0 + b_1 t)$ and $(t, b_2 + b_3 t)$, as this was believed to represent the geographical situation more precisely.

The results are shown in Figure 2.6. The axes are parameterised as $(t, 3.85 + 2.19t)$ with $k = 16.16$ and $\lambda = 3.097$ for the latitude, and $(t, -0.57 + 1.62t)$ with $k = 10.06$ and $\lambda = 0.15$ for the longitude. The loss values are 0.6310 and 0.5691 respectively, which show that these axes fit the data reasonably well (see Section 2.4 for more information about the goodness of fit).

Town	Latitude	Longitude
Birmingham	52.48	1.68
Bristol	51.52	2.58
Carlisle	54.62	3.15
Leeds	54.05	1.25
Liverpool	53.5	3.07
London	51.52	0.10
Manchester	53.33	2.15
Newcastle	54.98	1.60
Norwich	52.77	-1.35
Oxford	51.62	1.08

Table 2.1: Longitude and latitude values for selected towns and cities in the UK (www.theaa.com (2008))

It is possible to add anchor points to the axes. For one axis, by defining t_{MIN} as slightly lower than the value t_i from the object projecting lowest onto the axis, it is possible to calculate the coordinates of the minimum anchor. The length along the axis from the zero point to t_{MIN}, M_i say, is given by

$$M_i = \int_0^{t_{\text{MIN}}} \sqrt{x_1'(t)^2 + x_2'(t)^2}\text{dt} + k.$$

Thus the coordinates of t_{MIN} are given by $(x_1(M_i), x_2(M_i))$. Similarly, an anchor point can be described at the opposite end of the axes by defining t_{MAX} as slightly larger than the t_i from the object projecting highest onto the axis. In Figure 2.6, the object that projects lowest on the Latitude axis is London, with a t_i of 51.40, whilst the largest is Newcastle, with a t_i of 55.01. Similarly for the Longitude, the extreme points are Norwich ($t_i = -1.39$) and Carlisle ($t_i = 3.19$). Thus for the Latitude, $t_{\text{MIN}} = 51.3$ and $t_{\text{MAX}} = 55.1$, whilst for the Longitude $t_{\text{MIN}} = -1.40$ and $t_{\text{MAX}} = 3.20$. These result in the

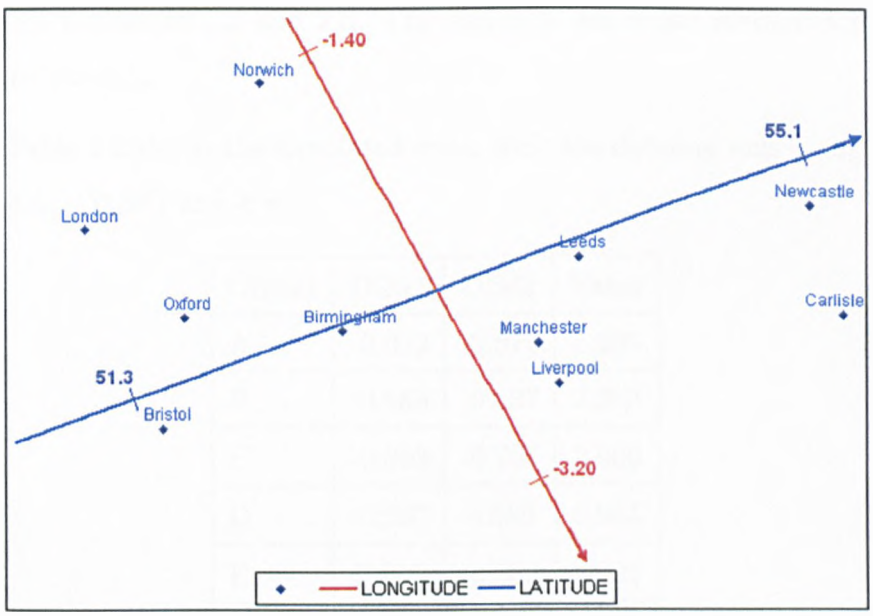

Figure 2.6: Axes for longitude and latitude added to MDS map

anchor points shown in Figure 2.6.

2.3.2 Simulated data

This section details the application of the methodology to simulated data. As the data has been user-generated, then it is known what the results should be. This gives a measure of how well the methodology works.

A 2-dimensional map was constructed showing 10 objects, with co-ordinates in both dimensions randomly selected from a uniform distribution between -1 and $+1$. To create the observed values, a quadratoc axis $(b_1 t, b_2 t + b_3 t^2)$ was chosen and the projection points from the objects to the axis calculated using the methodology in Section 2.2.1, as

$$l_i = \int_0^{t_i} \sqrt{(b_1)^2 + (b_2 + 2b_3 t)^2} \quad \mathrm{d}t + k.$$

based on Equations 2.5 and 2.6. The aim is to see if the methodology can recreate the axis.

Table 2.2 shows the simulated data, with the defining axis being $(1.1t, 0.4t - 0.3t^2)$ and $k = 1$.

Object	DIM1	DIM2	Value
A	-0.052	-0.671	1.395
B	-0.683	-0.537	2.323
C	-0.368	-0.707	2.000
D	-0.297	0.586	0.934
E	0.665	-0.532	0.021
F	-0.358	0.128	1.245
G	-0.872	0.631	1.545
H	0.527	0.460	0.110
I	-0.802	-0.071	2.007
J	-0.151	0.359	0.869

Table 2.2: Simulated data

As the values calculated are the actual projection points, then it is expected that the loss function will be zero, as $l_i = z_i$, from Equation 2.7.

The calculated parameters for the axis are $(1.091t, 0.397t - 0.295t^2)$, with $k = 0.977$ and a loss of 5.6611×10^{-6}. The loss is not exactly zero due to the limitations of the fitting software. The simulated and calculated results are shown in Figure 2.7.

As can be seen in Figure 2.7, the calculated axis matches the original simulated axis perfectly.

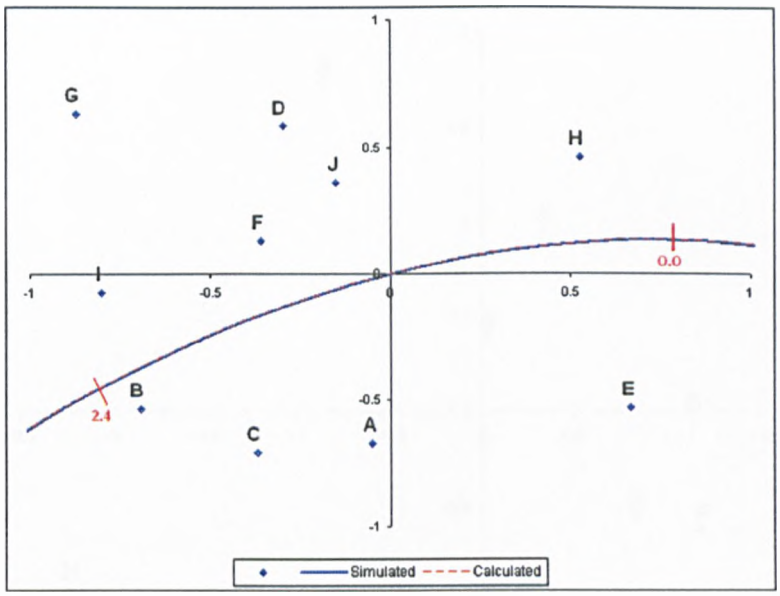

Figure 2.7: Fitted axis based on simulated data

2.3.3 Adding meaning to an MDS map produced from a consumer survey

In Section 1.7.2, a sensory test on deodorants was used to demonstrate the biplot. Here, the same data set is used to demonstrate the use of the curved axes. First, a non-metric MDS map is produced from the data, as seen in Figure 2.8. Then each attribute will be overlaid, using $(b_1 t, b_2 t + b_3 t^2)$ as the basis for the parameterisation of each axis.

The first attribute to be fitted is the Loudness of the Spray. The axis to describe this variable is given by $(0.528t, 6.924t + 1.159t^2)$, with a loss value of 4.87. As can be seen in Figure 2.9, this axis points from the bottom of the map towards the top - in other words products A and B have the loudest spray, whilst products C and H have the quietest. Looking at the raw data

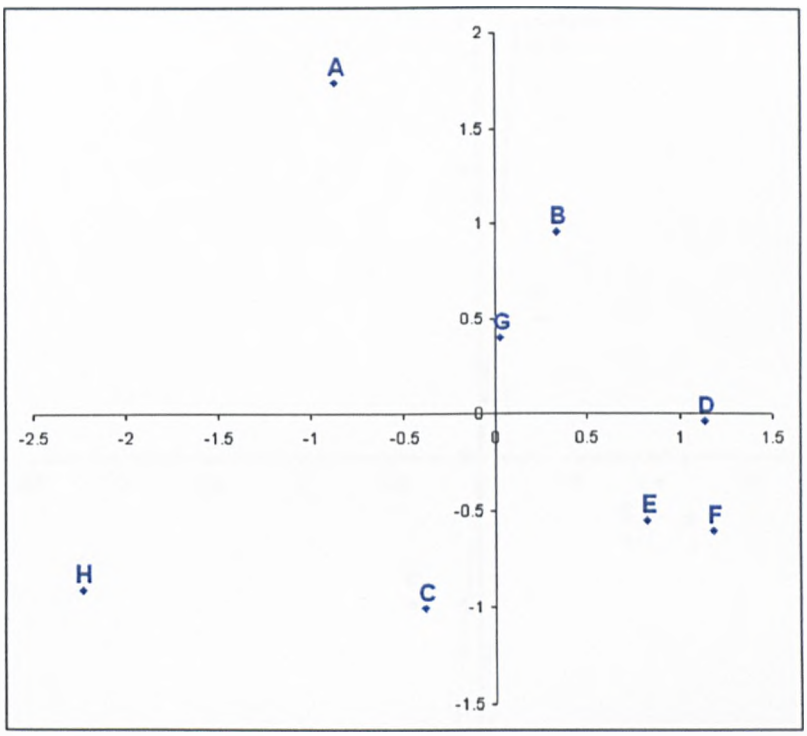

Figure 2.8: Non-metric MDS map of deodorant sensory data

in Table 1.3 and the biplot in Figure 1.5, it can be seen that this is indeed the case.

A second axis for Ease of use of spray has the parameters $(4.077t, -0.708t - 1.154t^2)$, and a loss of 0.395. Here, there is an improved fit over the biplot due to the quadratic nature of the axis. This axis goes from the left of the screen to the right, as shown in Figure 2.10, and again matches the biplot and the raw data.

A third axis for Evenness of spray is shown in Figure 2.11. It has parameters $(-1.663t, 1.153t - 4.353t^2)$, and a loss of 12.297. This is a highly quadratic axis with a large loss value, but a close investigation of the raw data shows the reason for this. First the products are not significantly different

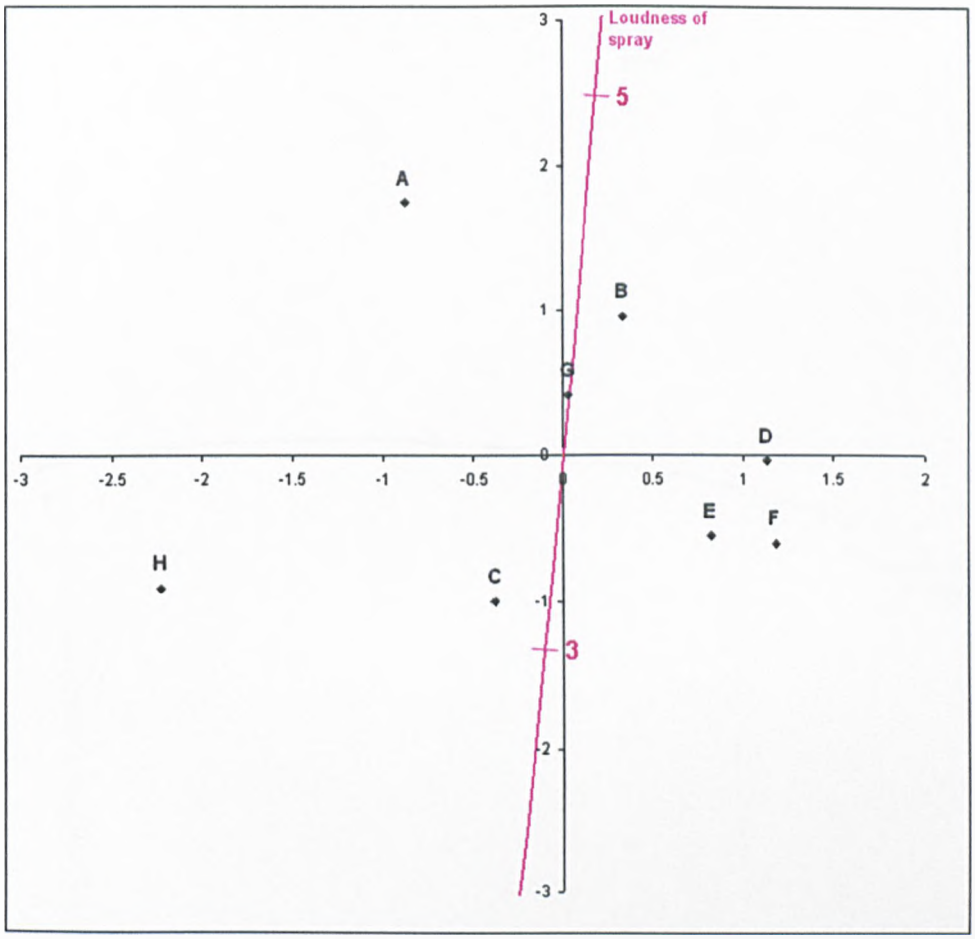

Figure 2.9: Non-metric MDS map of sensory data with axis for Loudness of spray overlaid

on this attribute - therefore differences are just due to noise. Second, if the products are ranked in order of their Evenness of spray, the sequence $A > G > D > F > E > C > H > B$ is obtained. Looking at the map, A and B are very close together, and so it is difficult to fit an axis for this attribute, and this is reflected in the high loss value. The fitted axis is the best representation of the data - this is not obvious from the biplot but is very obvious here.

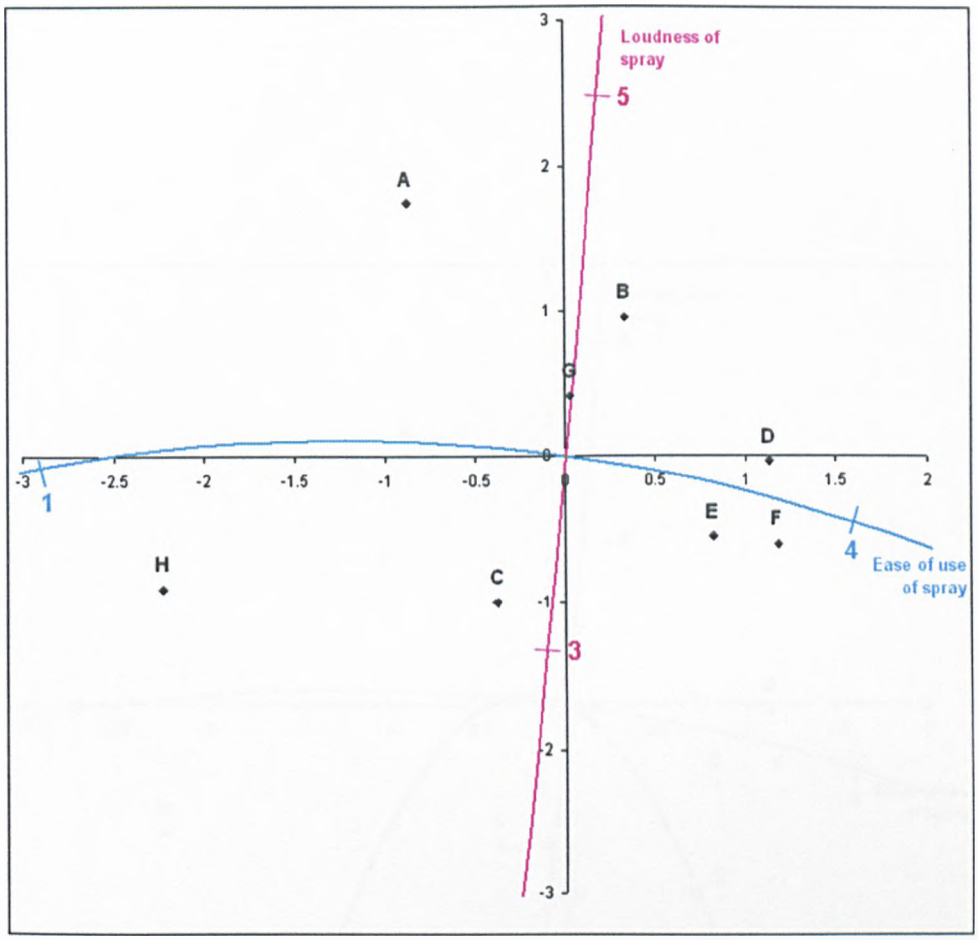

Figure 2.10: Non-metric MDS map of sensory data with axes for Loudness of spray and Ease of use overlaid

Repeating the process for each attribute results in the top plot in Figure 2.12. The second plot is the biplot repeated for comparison purposes.

Comparing the two plots shows that they are showing the same message. The products are in a similar place relative to each other, and each of the axes tends to point in the same direction. The curved axes allow for a better fitting map for the reasons discussed previously. The values for the parameters are shown in Table 2.3.

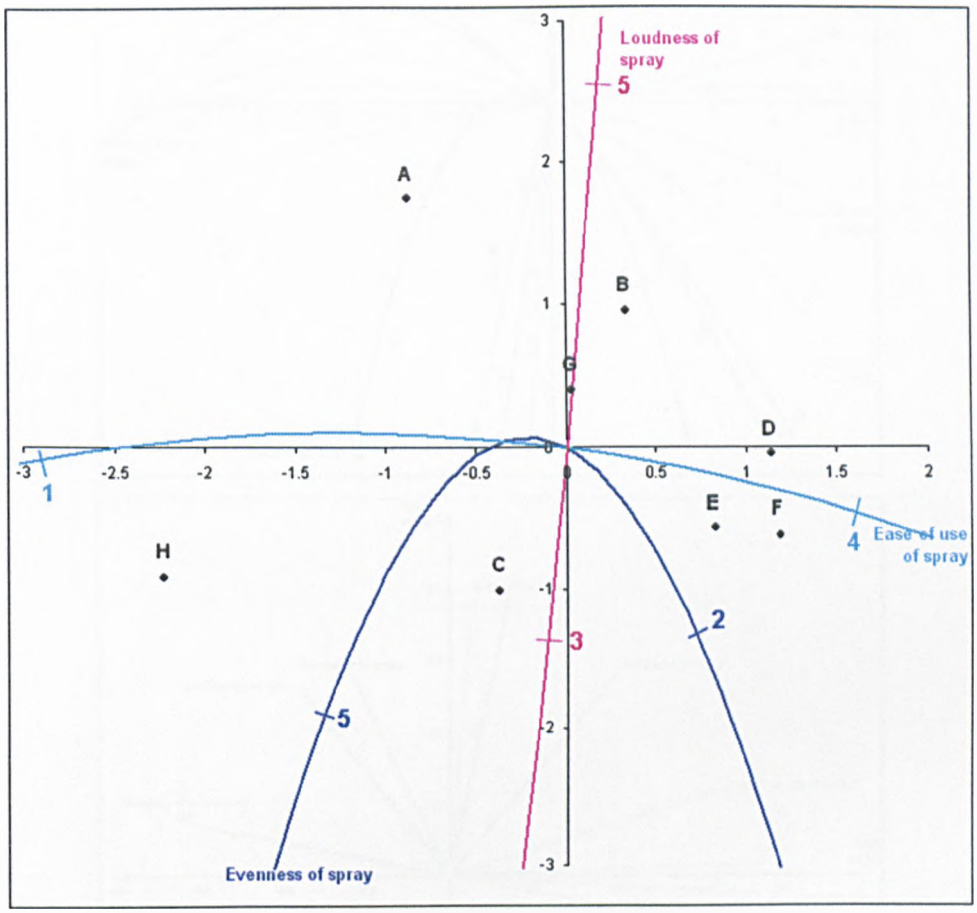

Figure 2.11: Non-metric MDS map of sensory data with three axes overlaid

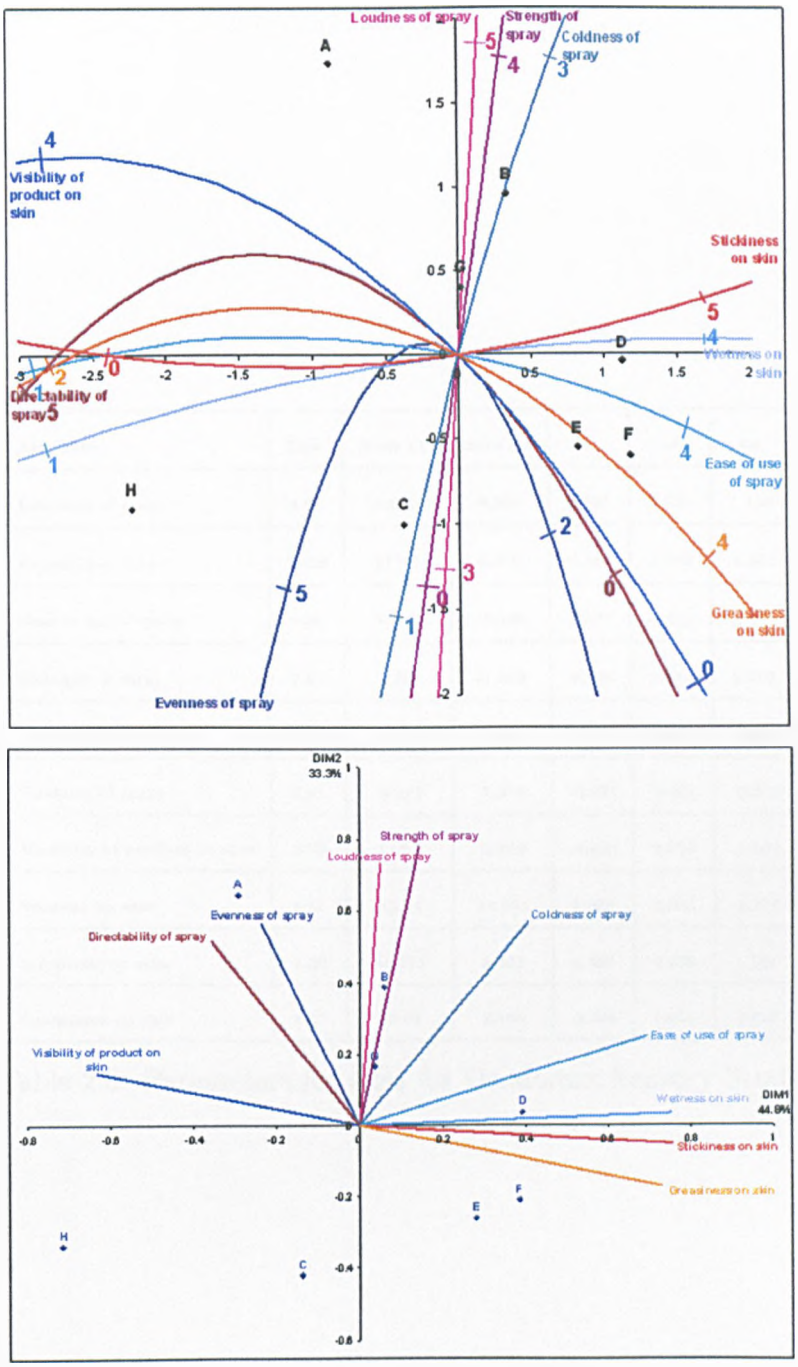

Figure 2.12: Curvlinear plot (top) and biplot (bottom) of deodorant sensory data

Attribute	Loss	Scale (λ)	Zero (k)	b1	b2	b3
Loudness of spray	4.87	4.620	5.692	0.528	6.924	1.159
Evenness of spray	12.29	271.6	0.370	-1.662	1.153	-4.353
Ease of use of spray	0.39	7.262	15.168	4.077	-0.708	-1.154
Strength of spray	2.67	0.232	-1.543	0.328	1.949	0.056
Directability of spray	0.89	11.994	6.780	2.287	-2.020	-1.680
Coldness of spray	6.61	8.920	7.975	-0.627	-2.036	-0.300
Visibility of product on skin	3.75	113.0	2.776	-4.093	3.726	-2.902
Wetness on skin	5.51	78.14	14.391	6.035	0.643	-1.171
Stickiness on skin	1.39	46.771	8.461	4.899	0.568	1.181
Greasiness on skin	2.87	16.01	2.408	-2.346	1.018	-0.902

Table 2.3: Parameters for axes for Deodorant Sensory Study

2.3.4 Adding meaning to the consumers

In Subsection 2.3.3, the technique was demonstrated on only 8 objects - the products in the study. However, there is no limit to the number of objects that the technique can use. In the study, 150 people were asked to score their overall opinion of each product. An MDS map can be produced on these people for each product in turn, and the technique used to overlay their overall opinion scores. This has been carried out in Figure 2.13 looking at product A.

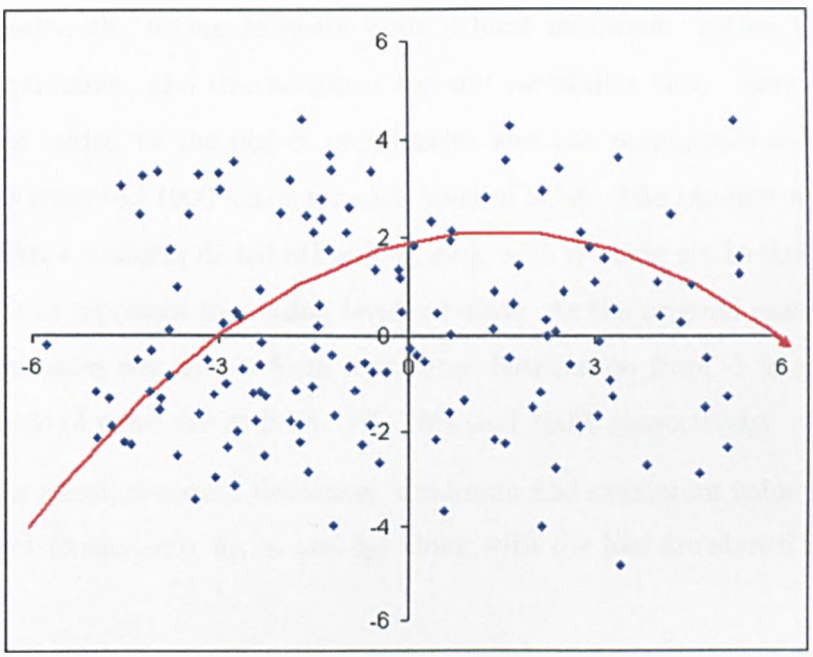

Figure 2.13: MDS map of consumers, with overall opinion axis overlaid

The axis in Figure 2.13 has parameters $(-3.2 + 6.3t, 6.1t - 4.3433t^2)$, and a loss of 20.81. It shows that there are differences between consumers, with those to the right hand side of the plot preferring the product to those on the left hand side. Similar patterns are seen for the other products.

2.4 Goodness of fit

Section 2.3 has shown that the methodology can prefectly recreate the curved axis for the simulated data. However, this is an exceptional example, in that most data contains noise which will affect the axis fitting. This section shows what happens as increasing levels of noise are added to the simulated data (to both the object coordinates, and the object scores).

The simulated data from Section 2.3.2 was used to show the effect of increasing noise. First 1000 repeats were run of the data with no noise. Occasionally, the fitting software finds a local minimum, rather than the overall minimum, and this accounts for any variability here. Next random noise was added to the object coordinates and the parameters calculated. This was repeated 1000 times for each level of noise. The random noise was taken from a uniform distribution $[-n, +n]$, with n being set to 0.001, 0.01, 0.1 and 1 to represent increasing levels of noise. As the original raw data for the coordinates was drawn form a uniform distribution from -1 to +1, then these levels of noise are at 0.1%, 1%, 10% and 100% respectively.

The mean, standard deviation, minimum and maximum value for each parameter (scale, zero, b_1, b_2 and b_3) along with the loss are shown in Table 2.4.

0%	Loss	Scale	Zero	b_1	b_2	b_3
Mean	5.687×10^{-6}	0.977	-0.0136	1.0906	0.397	-0.295
Std.Dev.	2.637×10^{-7}	0.00576	0.0109	0.00283	0.000933	0.001
Max.	5.66×10^{-6}	0.966	-0.0031	1.0977	0.400	-0.281
Min.	8.31×10^{-6}	0.919	-0.125	1.0624	0.387	-0.299
0.1%	Loss	Scale	Zero	b_1	b_2	b_3
Mean	1.487×10^{-5}	0.975	-0.169	1.089	0.396	-0.294
Std Dev	0.00005	0.033	0.030	0.016	0.005	0.008
Max	0.00530	1.008	0.011	1.107	0.406	-0.216
Min	1.36×10^{-6}	0.663	-0.266	0.933	0.346	-0.301
1%	Loss	Scale	Zero	b_1	b_2	b_3
Mean	0.00026	0.960	-0.035	1.082	0.397	-0.290
Std Dev	0.00142	0.082	0.090	0.038	0.014	0.019
Max	0.01430	1.160	0.076	1.163	0.469	-0.237
Min	4.26×10^{-6}	0.747	-0.621	0.974	0.361	-0.336
10%	Loss	Scale	Zero	b_1	b_2	b_3
Mean	0.01400	0.917	-0.127	0.934	0.361	-0.254
Std Dev	0.02750	0.602	0.590	0.598	0.259	0.146
Max	0.29100	3.078	3.787	1.689	0.618	0.405
Min	0.00075	0.018	-1.175	-4.492	-2.087	-0.621
100%	Loss	Scale	Zero	b_1	b_2	b_3
Mean	0.91000	3.717	3.513	-0.686	-0.696	-4.610
Std Dev	0.50200	9.234	9.122	3.771	2.988	27.530
Mean	2.26840	61.917	60.950	3.890	2.260	6.400
Min	0.14500	0.016	-6.510	-31.370	-20.230	-285.800

Table 2.4: Summary data from adding noise to coordinates

The distributions of the parameters are seen in Figures 2.14 to 2.18. Looking at Table 2.4 it can be seen that as the noise increases, the loss value increases. This is to be expected, as it is harder to fit the axis to the values. In addition, the b-parameters get further away from their true values of ($b_1 = 1.1, b_2 = 0.4$ and $b_3 = -0.3$), along with an increase in the standard deviation. However, the mean values are still fairly close to the expected values, and axes produced using these parameter values still match the expected axes. In addition, when looking at the parameter distributions, they are skewed towards the expected values, and it is only a relatively few values that are lying away from the expected.

At 100% noise, then it appears that the values produced are very unreliable. However, looking at the distributions in Figure 2.18, then there are about 10 values out of the 1000 repeats that can be thought of as extreme outliers. If these values are removed, then the mean scores are as expected.

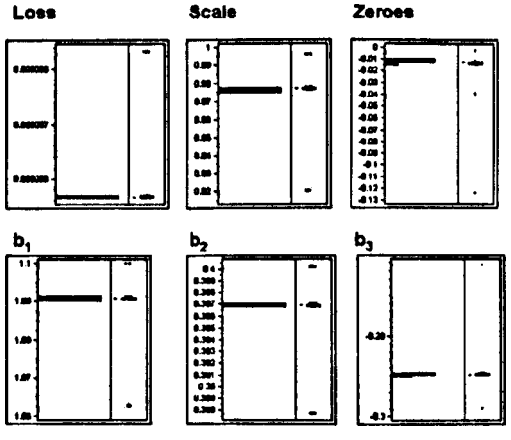

Figure 2.14: Distribution of parameters when there is no noise

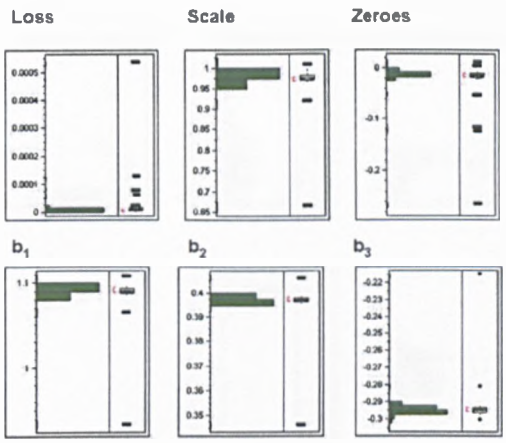

Figure 2.15: Distribution of parameters when there is 0.1% noise in the coordinates

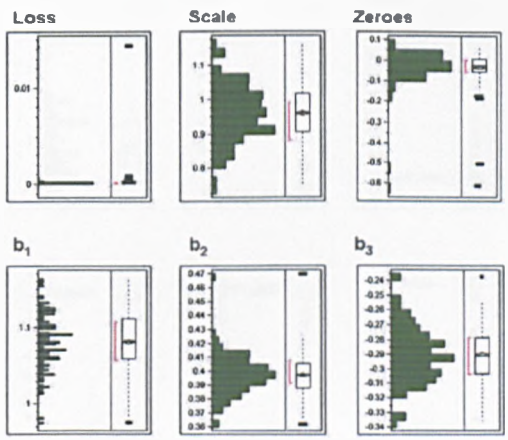

Figure 2.16: Distribution of parameters when there is 1% noise in the coordinates

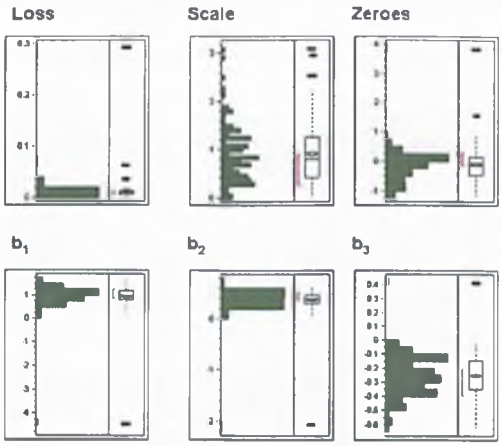

Figure 2.17: Distribution of parameters when there is 10% noise in the coordinates

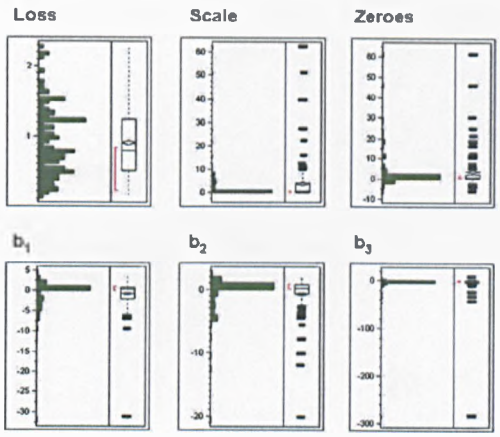

Figure 2.18: Distribution of parameters when there is 100% noise in the coordinates

The procedure was repeated but with noise being added to the observed values of the attribute (y_i in Equation 2.7) instead of the object coordinates. Results are shown in Table 2.5 and Figures 2.19 to 2.22.

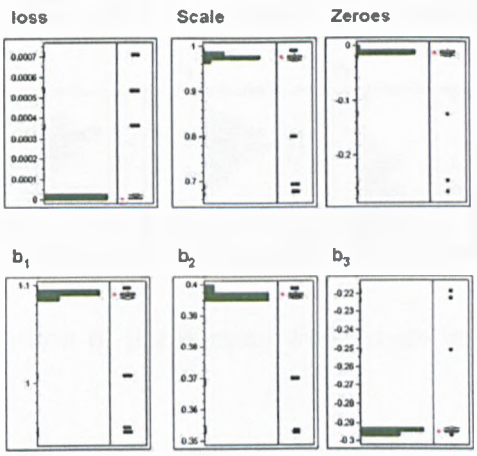

Figure 2.19: Distribution of parameters when there is 0.1% noise in the attribute scores

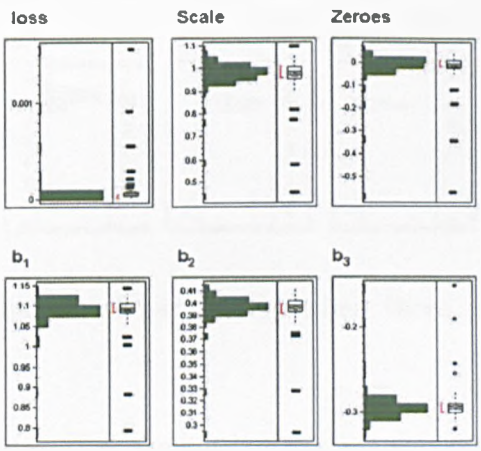

Figure 2.20: Distribution of parameters when there is 1% noise in the attribute scores

The fact that despite increasing levels of noise, the calculated parame-

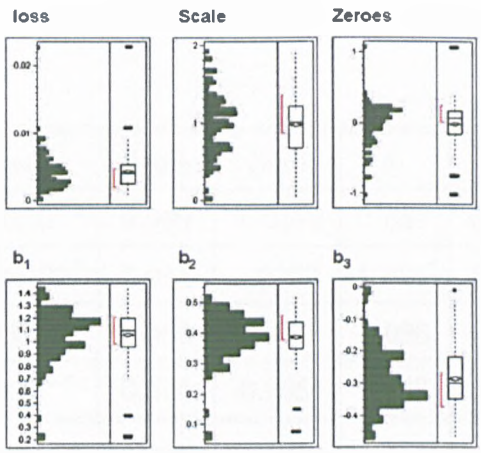

Figure 2.21: Distribution of parameters when there is 10% noise in the attribute scores

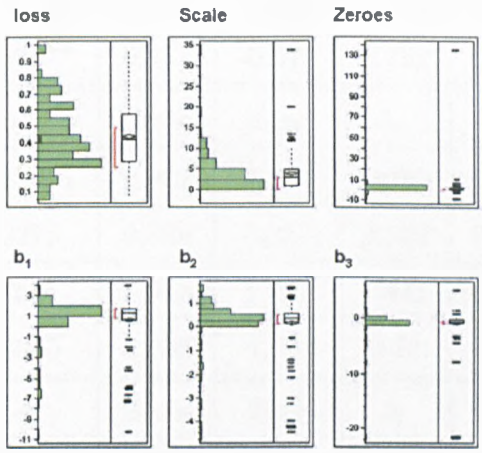

Figure 2.22: Distribution of parameters when there is 100% noise in the attribute scores

ters are consistent shows that the methodology is accurately capable of fitting the data.

0.1%	Loss	Scale	Zero	b_1	b_2	b_3
Mean	2.125×10^{-5}	0.969	-0.0179	1.087	0.396	-0.293
SD	9.299×10^{-5}	0.044	0.0363	0.021	0.007	0.012
Max	0.00071	0.988	-0.0056	1.096	0.399	-0.219
Min	2.82×10^{-6}	0.674	-0.2686	0.948	0.352	-0.297
1%	Loss	Scale	Zero	b_1	b_2	b_3
Mean	0.0000757	0.971	-0.017	1.087	0.396	-0.293
SD	0.000173	0.076	0.069	0.039	0.013	0.019
Max	0.00155	1.092	0.054	1.139	0.413	-0.153
Min	8.33×10^{-6}	0.447	-0.576	0.791	0.293	-0.321
10%	Loss	Scale	Zero	b_1	b_2	b_3
Mean	0.0043391	0.982	-0.014	1.06	0.385	-0.286
SD	0.00275	0.389	0.287	0.209	0.075	0.093
Max	0.0227	1.917	1.051	1.441	0.524	-0.012
Min	0.00035	0.033	-1.037	0.21	0.074	-0.473
100%	Loss	Scale	Zero	b_1	b_2	b_3
Mean	0.436	3.908	2.091	0.626	0.139	-0.826
SD	0.197	4.439	12.698	2.316	1.047	2.315
Max	0.967	33.541	134	3.82	1.592	3.97
Min	0.0552	0.009	-9.7	-10.36	-4.535	-21.93

Table 2.5: Summary data from adding noise to attribute scores

2.5 Model selection

So far only linear and quadratic functions have been used to define the axes. However, in theory, any differentiable function could be used, as long as the derivatives are not zero. This leads to the question - what is the best fitting axis? This can be answered using one of several model selection methods.

Model selection is the task of choosing a model from a set of potential models with the best inductive bias, which in practice means selecting parameters in an attempt to create a model of optimal complexity given (finite) training data. A classical example is the principle of Occam's Razor, which assumes that the simplest consistent hypothesis about the target function is actually the best. Thus model selection is a bias versus variance trade-off, and this is the statistical process of parsimony. Inference under models with too few parameters can be biased, whilst with models that have too many parameters, there may be poor precision, or identification of effects that are, in effect, spurious. These considerations call for a balance between under- and over- fitted models - the so called *model selection problem* (Forster (2000) and Burnham and Anderson (2004a)).

Here, cross-validation will be used to demonstrate model selection on the simulated axis of Section 2.3.2. Cross-validation is a method of evaluating given models by means of their forecasts, and to choose a model with proper complexity (Hjorth (1982)). In theory, cross-validation is the statistical practice of partitioning a sample of data into subsets such that the analysis is initially performed on a single subset (the training set), while the other subset(s) are retained for subsequent use in confirming and validating the initial analysis (the validation or testing sets).

Different types of cross-validation exist, depending on how the parti-

tioning occurs:

K-fold cross-validation partitions the original sample into K subsamples. Of the K subsamples, a single subsample is retained as the validation data for testing the model, and the remaining $K-1$ subsamples are used as the training data. The cross-validation process is then repeated K times, so that each one of the subsamples is used once as the validation data. The K results are then combined (often by averaging) to produce a single estimation.

Leave-one-out-cross-validation is the same as K-fold cross-validation, but with K being equal to the number of observations in the original sample - thus each validation set consists of one observation.

After carrying out the cross-validation, the parameter estimation error can be computed. Common error metrics are the mean squared error and the root mean squared error, respecitvely giving the estimated variance and standard deviation of the cross-validation. If this process is repeated for each possible model, the best-fitting model can be selected as that with the lowest error.

In terms of axis fitting, each object can represent a subsample for the leave-one-out-cross-validation. An object can be removed from the sample, and the remaining objects used to fit the axis. Once the parameters have been calculated, then the score of the removed object can be calculated by:

$$Score'_i = \lambda \left[\int_0^{t_i} \sqrt{x'_1(t)^2 + x'_2(t)^2} \mathrm{dt} + k \right]$$

where $Score'_i$ is the estimated score for object i calculated when it has been removed from the data set.

An error term for each object can then be calculated by

$$Error_i = (Score_i - Score'_i)$$

The MSE and RMSE can then be calculated. Such an approach was carried out on several possible axes for the simulated data in Section 2.3.2. The results can be found in Table 2.6.

Model	MSE	RMSE
$(b_1 t, b_2 t)$	0.7053	0.8398
$(b_1 t, b_2 t + b_3 t^2)$	0.0024	0.0492
$(b_1 t + b_2 t^2, b_3 t + b_4 t^2)$	1.7254	1.3135

Table 2.6: Results of LOOCV on axis selection for simulated data

The model with the lowest MSE and RMSE is the second one $(b_1 t, b_2 t + b_3 t^2)$, so it is this axis that describes the data best. This is not surprising, as this is the equation used in the simulation of the data.

Using the data from Section 2.3.3, the leave-one-out-cross-validation can be used to find the best-fitting axis for loudness of spray. The statistics for the fit of several models are shown in Table 2.7.

Model	MSE	RMSE
$(b_1 t, b_2 t)$	12.476	1.249
$(b_1 t, b_2 t + b_3 t^2)$	93.165	3.412
$(b_1 t + b_2 t^2, b_3 t + b_4 t^2)$	191.921	4.896

Table 2.7: Results of LOOCV on axis selection for loudness of spray data

The model with the lowest error is $(b_1 t, b_2 t)$. This is not surprising as close inspection of Figure 2.9 shows that when the axis was fitted previously, the quadratic term was close to zero. Fitting the 'best' axis gives parameters of $(0.719t, 8.596t)$, with a loss of 23.658, zero point of 17.087 and scale factor of 4.0066.

Other model selection methods can be used. For example,

- Akaike Information Criterion (AIC)
- Bayesian Information Criterion (BIC)
- Structural Risk Minimisation with VC-dimensions (SRMVC)

These three methods have the advantage over cross-validation in that only the training error is needed. Asymptotically though, AIC and leave-one-out-cross-validation should be identical (McQuarrie and Tsai (1998)).

2.6 Extensions to n-dimensions

So far, this chapter has dealt with fitting axes to 2-dimensional MDS maps. The methodology, however, is extendable to p-dimensions.

Let $y = \{y_r(t)\}$ $r = 1, ..., p$ be the axis to be fitted. Now P_i, the projection point of object i onto the axis can be given by

$$\sum_{r=1}^{p} y_r'(P_i)\left[y_r(P_i) - x_{ir}\right] = 0$$

where x_{ir} is the coordinate for object i in dimensions r.

Once the P_i has been calculated, the length along the axis from this projection point to the origin is given by

$$l_i = \int_0^{t_i} \sqrt{\sum_{r=1}^{p} y_r'(t)^2} \ \mathrm{dt} + k.$$

The goodness-of-fit measure for the p-dimensional axis is thus

$$\text{goodness of fit} = \frac{\sum_{i=1}^{n}\left[z_i - (l_i + k)\,\lambda\right]^2}{n} \tag{2.8}$$

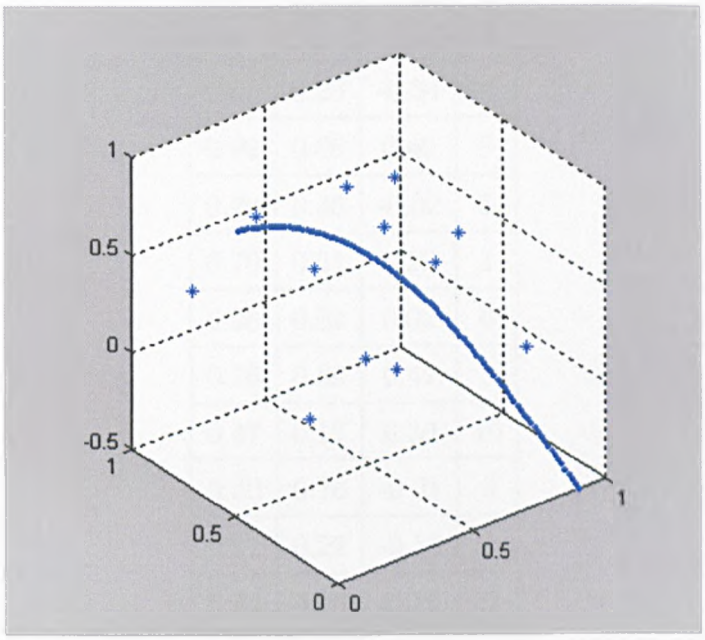

Figure 2.23: A 3-dimensional MDS plot with curved axes

Figure 2.23 shows an example of a 3-dimensional MDS plot, with an overlaid axis. The data used to generate this plot can be found in Table 2.8.

The equations for the axis in Figure 2.9 are

$$(0.98\sin(t), 1.02\cos(t), 0.54t)$$

with a goodness-of-fit of 3.244.

The discussions about model selection and goodness-of-fit in Sections 2.4 and 2.5 are applicable here too.

i_1	i_2	i_3	Z
0.63	0.23	-0.34	9
0.92	0.66	0.40	9
0.23	0.36	-0.02	5
0.70	0.51	0.28	1
0.98	0.81	0.02	6
0.76	0.69	0.42	3
0.47	0.17	0.30	10
0.60	0.76	-0.21	6
0.51	0.22	-0.18	4
0.44	0.16	-0.16	5
0.17	0.51	0.48	8
0.87	0.13	-0.37	7

Table 2.8: Simulated data for 3 dimensional MDS plot and axis to be fitted

2.7 Points for discussion

2.7.1 Measures of curvature

In vector calculus, the Frenet-Serret formulae describe the kinematic properties of a particle which moves along a continuous, differentiable curve in three-dimensional Euclidean space R^3 (Guggenheimer (1977)). In more detail, the formulae actually describe the derivatives of the tangent, normal and binormal unit vectors, in terms of each other (Struik (1961)). These ideas could be used as an alternative mechanism for describing the curves overlaid onto a 3-dimensional MDS map.

Suppose $\mathbf{r}(t)$ is a curve in Euclidean space which represents the axis as a

function of time (as t_i is used in previous sections). The Frenet-Serret formula can be applied to such curves if they are non-degenerate. This roughly means that they have curvature, but can be expressed more formally as the velocity vector $\mathbf{r}'(t)$ and the acceleration vector $\mathbf{r}''(t)$ not being perpendicular (Spivak (1999)).

Let $s(t)$ represent the arc length along the curve. The quantity s is used to give a natural parameterisation to the curve traced out by the trajectory over time, since many different paths may trace out the same geometrical curve by traversering it at different rates. In more detail, s is given by

$$s(t) = \int_0^t ||\mathbf{r}'(\tau)||d\tau$$

In addition, due to the assumption that $\mathbf{r}' \neq 0$, it is possible to solve for t as a function of s, and thus to write $\mathbf{r}(s) = \mathbf{r}(t(s))$ (Iyer and Vishveshwara (1993) and Crenshaw and Edelstein-Keshet (1993)). The curve is therefore parameterised in a preferred manner by its arc length.

With a non-degenerative curve $\mathbf{r}(s)$, parameterised by its arclength, the Frenet-Serret formulae can now be defined:

$$\text{The tangent vector } \mathbf{T} = \frac{d\mathbf{r}}{ds}. \tag{2.9}$$

$$\text{The normal vector } \mathbf{N} = \frac{\frac{d\mathbf{T}}{ds}}{\| \frac{d\mathbf{T}}{ds} \|}. \tag{2.10}$$

$$\text{The binormal unit vector } \mathbf{B} = \mathbf{T}\text{x}\mathbf{N}. \tag{2.11}$$

The binormal unit vector is defined as the cross-product of $\mathbf{T}$ and $\mathbf{N}$.

From equation 2.10 it follows that, since $\mathbf{T}$ always has to have unit magnitude, then $\mathbf{N}$ is always perpendicular to $\mathbf{T}$, whilst from equation 2.11, $\mathbf{B}$ is always perpendicular to $\mathbf{T}$ and $\mathbf{N}$. In other words, the three unit vectors

are all perpendicular to each other, and can be used to define a unique line through Euclidean space.

The Frenet-Serret formulae can thus now be given as

$$\begin{aligned}\frac{d\mathbf{T}}{ds} &= \kappa\mathbf{N} \\ \frac{d\mathbf{N}}{ds} &= -\kappa\mathbf{T} + \tau\mathbf{B} \\ \frac{d\mathbf{B}}{ds} &= -\tau\mathbf{N}\end{aligned} \tag{2.12}$$

where κ is the curvature and τ is the torsion.

The Frenet-Serret formulae are also known as the Frenet-Serret thereom, and can be more concisely represented using matrix notation. This gives a skew-symmetric matix.

$$\begin{bmatrix} \mathbf{T}' \\ \mathbf{N}' \\ \mathbf{B}' \end{bmatrix} = \begin{bmatrix} 0 & \kappa & 0 \\ -\kappa & 0 & \tau \\ 0 & -\tau & 0 \end{bmatrix} \begin{bmatrix} \mathbf{T} \\ \mathbf{N} \\ \mathbf{B} \end{bmatrix}$$

The Frenet-Serret formulae can be generalised to higher-dimensional Euclidean spaces - see Guggenheimer (1977) for more details.

Example 1

A simple example is demonstrated here in 2-dimensions, where only curvature is defined. For a curve $(x_1(t), x_2(t))$, the curvature is given by

$$\kappa = \frac{x_1'(t)\, x_2''(t) - x_2'(t)\, x_1''(t)}{\left(x_1'(t)^2 + x_2'(t)^2\right)^{\frac{3}{2}}}.$$

Taking the axis from Section 2.3.2, $(b_1 t, b_2 t + b_3 t^2)$, the curvature is

$$\kappa = \frac{b_1(2b_3) - (b_2 + 2b_3 t)\,0}{\left(b_1^2 + (b_2 + 2b_3 t)^2\right)^{\frac{3}{2}}},$$

$$\kappa = \frac{2b_1 b_3}{\left(b_1^2 + (b_2 + 2b_3 t)^2\right)^{\frac{3}{2}}}.$$

So with the axis parameterised as $b_1 = 1.091$, $b_2 = 0.397$, and $b_3 = -0.295$, the curvature is

$$\kappa = \frac{0.643}{\left(1.190 + (0.397 - 0.590t)^2\right)^{\frac{3}{2}}}.$$

Example 2

In three dimensions, let an example curve be defined as a helix $\mathbf{r} = (b\cos t, b\sin t, t)$. From equations 2.9 to 2.11 it is possible to calculate the Frenet-Serret formulae as

$$\mathbf{T} = \mathbf{r}' = \left(-\frac{b}{\sqrt{b^2+1}}\sin\frac{s}{\sqrt{b^2+1}}, \frac{b}{\sqrt{b^2+1}}\cos\frac{s}{\sqrt{b^2+1}}, \frac{1}{\sqrt{b^2+1}}\right),$$

$$\mathbf{N} = \frac{\mathbf{T}'}{\| \mathbf{T}' \|} = \left(-\cos\frac{s}{\sqrt{b^2+1}}, -\sin\frac{s}{\sqrt{b^2+1}}, 0\right),$$

and

$$\mathbf{B} = \mathbf{T} \times \mathbf{N} = \left(\frac{1}{\sqrt{b^2+1}}\sin\frac{s}{\sqrt{b^2+1}}, -\frac{1}{\sqrt{b^2+1}}\cos\frac{s}{\sqrt{b^2+1}}, \frac{b}{\sqrt{b^2+1}}\right),$$

where $s = t\sqrt{b^2+1}$.

Using Equation 2.12 the curvature and torsion can be calculated,

$$\kappa = \frac{b}{b^2+1}$$

and

$$\tau = \frac{1}{b^2+1}.$$

2.7.2 Curvilinear Coordinates

If p curved axes, representing some variables, are fitted to an MDS plot which is in p-dimensions, then these p axes can be considered as curvilinear axes defining the MDS space. The points in the space, which represent the

products, can then be defined by these new axes. For example, in Figure 2.13, if, in addition to the overall opinion axis, another axis was defined by likelihood to buy, then overall opinion and likelihood to buy could be used as curvilinear coordinates. The plot could then be compared to the simple scatterplot of likelihood to buy against overall opinion, in an attempt to see how the multivariate attributes affect these two scores.

Chapter 3

An MDS approach to visualising multivariate paired comparison data

3.1 Introduction

The theory and practice of paired comparison testing have advanced considerably since Thurstone formulated his law of comparative judgment in 1927. The law of comparative judgment (more correctly, the model of comparative judgment) is a mathematical representation of a discriminal process. This is any process in which a comparison is made between pairs of a collection of entities, with respect to magnitudes of an attribute, trait, attitude, and so on. Thurstone introduced the concept of an underlying psychological scale (or latent variable) on which the comparison is made.

The law indicates that the scale difference between any two stimuli is a random variable whose probability density function forms a normal distri-

bution. Thus, if π_i is the pyschological scale value of stimulus i, then

$$\pi_i - \pi_j = z_{ij}\sqrt{\sigma_i^2 + \sigma_j^2 - 2\rho_{ij}\sigma_i\sigma_j}$$

where z_{ij} is the normal deviate corresponding to the proportion of times stimulus i is chosen over stimulus j, as the number of comparisons made tends to infinity; σ_i and σ_j are the standard deviations for stimuli i and j respectively; and ρ_{ij} is the correlation coefficient between stimuli i and j.

In the fifth case of Thurstone's Law of Comparative Judgement, all stimuli share the same standard deviation among observations, i.e. $\sigma_i = \sigma_j = \sigma$, and the correlation between each stimulus is zero. Accordingly, the scale value difference, $\pi_i - \pi_j$, has a standard deviation $\left(\sigma\sqrt{2}\right)$, and thus $\pi_i - \pi_j = z_{ij}\sigma\sqrt{2}$ (Cui (2001)).

Since its early beginnings, the paired comparison test (or, more formally, the 2-Alternative Forced Choice test) has become one of the most widely used of sensory tests (McBride et al. (1984), Hymann and Lawless (1999), and Anon (2005)). Paired comparison testing generally nowadays refers to any process of comparing objects (or stimuli) in pairs to judge which of each pair is preferred, or has a greater amount of some quantitative property (Marden (1996)). This type of experiment on a set of t objects results in $\binom{t}{2}$ paired comparisons, usually shown to panellists in a random order. Traditionally a set number of panellists are asked to carry out all the $\binom{t}{2}$ comparisons. However, this is not always necessary, especially when there is a large number of objects.

Bradley and Terry (1952) introduced a simple and appealing model for the analysis of such tests, when t objects are compared in pairs, with object rating parameters, $\pi_1, \ldots \pi_t$.

In Bradley (1955), large-sample results and the asymptotic distribution of the maximum-likelihood estimators are given, whilst Ford (1957) described an iterative solution of the likelihood equations. Hunter (2004) builds on a theory of algorithms known by the initials of MM, for minorization-maximization, and presents a powerful technique for producing iterative maximum likelihood estimation algorithms for a wide class of generalisations of the Bradley-Terry model.

The basic Bradley-Terry model has been extensively discussed in the literature (David (1988)) and various extensions have been proposed. To name just a few of these: ties (Rao and Kupper (1967), Davidson (1970), Kousgaard (1976)), order effects (Davidson and Beaver (1977), Fienberg (1979)), the incorporation of explanatory vaiables (Kousgaard (1984), Matthews and Morris (1995), Francis et al. (2002)), and ordinal paired comparison models (Agresti (1992), Boeckenholt and Dillon (1997)). The main importance here for sensory testing has been work of Dittrich et al. (1998) to include panellist effects.

When several attributes are being considered concurrently, then the Bradley-Terry model can be fitted for each one in turn. However, a multivariate approach might be preferable. This chapter first discusses the Bradley-Terry model in more detail, then describes a multivariate approach, leading to a map similar to the biplot, which shows how each object is scored on each attribute.

3.2 The Bradley-Terry model

The Bradley-Terry model (Bradley and Terry (1952) and Luce (1959)) is often applied to pairwise comparison data to scale preferences.

For the univariate case, consider t treatments, or objects, to be compared in pairs with treatment rating parameters $\pi_1, \ldots, \pi_t$. The Bradley-Terry model postulates that treatments have true ratings on a particular subjective continuum such that $\pi_i \geq 0$ and $\sum \pi_i = 1$. When treatment i is compared to treatment j, the probability P_{ij} that treatment i is ranked over treatment j (or receives a rank of 1) is given by

$$P_{ij} = \frac{\pi_i}{\pi_i + \pi_j} \tag{3.1}$$

for each and every pair of treatments. The model could alternatively be expressed as

$$\text{logit}\,(P_{ij}) = \mu_i - \mu_j$$

where $\mu_i = \log \pi_i$ for all i.

Define $n_{ijk} = 1$ if, in the kth paired comparison of treatments i and j, treatment i is selected over treatment j, and $n_{ijk} = 0$ otherwise. Let treatments i and j be compared N_{ij} independent times. Then $n_{ij} = \sum_k n_{ijk}$ is the number of comparisons out of N_{ij} when the ith treatment is selected or preferred over the jth treatment. If the rating parameters remain the same from comparison to comparison, then n_{ij} follows a binomial distribution. The likelihood of the total experiment is the product of $t\,(t-1)\,/2$ binomial functions (Bradley and El-Helbawy (1976)). Maximum likelihood estimates $\hat{\pi}_i$ of π_i are obtained by solving the likelihood function iteratively. The likelihood equations are

$$\hat{\pi}_i = \frac{n_i}{\sum_{j \neq i} N_{ij}\,(\hat{\pi}_i + \hat{\pi}_j)^{-1}} \quad , i = 1, \ldots, t, \tag{3.2}$$

where $n_i = \sum_{j \neq i} n_{ij}$. This equation can be fitted using any number of standard statistical packages. For the purpose of this thesis, the PROC LOGISTIC routine in SAS was used, which had the benefit of allowing additional effects, such as panellist, to be fitted.

Table 3.1 shows the results of a paired comparison test with five products where each pair was compared eighteen times.

	A	B	C	D	E
A	.	11	9	7	3
B	7	.	11	9	7
C	9	7	.	8	6
D	11	9	10	.	6
E	15	11	12	12	.

Table 3.1: Results from a paired comparison experiment. The values represent the number of times the row object was chosen over the column product.

The products preferred are in the rows, so for example product A was chosen over B eleven times. The results of the Bradley-Terry model give treatment rating parameters of $\hat{\pi}_A = 0.14$, $\hat{\pi}_B = 0.17$, $\hat{\pi}_C = 0.14$, $\hat{\pi}_D = 0.19$, and $\hat{\pi}_E = 0.36$. A plot of these values on an axis can be seen in Figure 3.1.

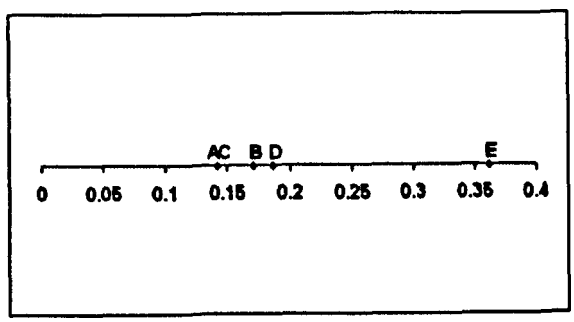

Figure 3.1: Treatment rating parameters from the univariate Bradley-Terry model

It can be seen that product E is much preferred over the other products, where for instance, the probability that product E is preferred over product

D is estimated as

$$\begin{aligned}\hat{P}_{ED} &= \frac{\hat{\pi}_E}{\hat{\pi}_E + \hat{\pi}_D} \\ &= \frac{0.36}{0.36 + 0.17} \\ &= 0.68\end{aligned}$$

Causeur and Husson (2005) introduced a 2-dimensional extension of the model that accounts for interactions between the compared objects, building on the work of Hunter (2004). This allowed for plotting of the treatment parameters when they are not transitively related. For instance, in a sensory context, it can be observed that the product A, say, is markedly better than B, B is also markedly better than C; however, C is preferred to A. This work though is still univariate in nature, dealing with comparisons on one attribute only.

The hypothesis of interest in sensory testing is that all objects are equally preferred or are of equal ratings. In terms of the Bradley-Terry model parameters the hypothesis that objects have equal ratings can be expressed as:

$H_0 : \pi_i = 1/t$ for all i;

$H_1 : \pi_i \neq 1/t$ for some i.

For testing the null hypothesis of equal ratings, the likelihood ratio test turns out to depend on the statistic

$$B = \sum_{i<j} N_{ij} \ln(\hat{\pi}_i + \hat{\pi}_j) - \sum n_i \ln \hat{\pi}_i,$$

which, for a large scale experiment has an approximate distribution given by

$$(1.3863)N - 2B \approx \chi^2_{(t-1)},$$

where N is the total number of all paired comparisons in the experiment (Sen and Puri (1967)).

For the example data shown in Table 3.1, the null hypothesis of no difference between objects is strongly rejected. Following the calculations, $B = 92.96$, so the test statistic equals 62.20, which when compared with a χ^2_4 distribution gives a p-value of $9.59e^{-13}$. Thus there is at least one significant difference between the objects. To highlight the nature of these differences the multiple-comparison test is based on finding confidence intervals for $(\pi_i - \pi_j)$, say, which are given by

$$(\hat{\pi}_i - \hat{\pi}_j) \pm z_{\alpha/2}\sqrt{\text{var}\,(\hat{\pi}_i - \hat{\pi}_j)}$$

where the variance of $(\hat{\pi}_i - \hat{\pi}_j)$ is defined in terms of the variances and covariances of $\hat{\pi}_i$ and $\hat{\pi}_j$ (Bradley (1955) and Dykstra (1960)). Davidson (1970) established the large-sample joint distribution of $\hat{\pi}_i$, and that $(\hat{\pi}_i - \pi_i)$ have jointly an approximate multivariate normal distribution. These definitions allow the calculation of the required variances and covariances. David (1988) found that if all N_{ij} are equal to N, and when the null hypothesis of $\pi_i = 1/t$ is true, it follows that

$$\sigma_{ii} = 4\,(t-1)\,/Nt^4 \quad i = 1,\ldots,t,$$
$$\sigma_{ij} = -4/Nt^4 \quad i \neq j = 1,\ldots,t.$$

where σ_{ij} is the covariance between π_i and π_j, and σ_{ii} is the variance of π_i. Thus

$$\text{var}\,(\pi_i - \pi_j) = \sigma_{ii} + \sigma_{jj} - 2\sigma_{ij}.$$

This of course assumes independence of the comparisons, which is not strictly correct. In sensory testing, the panellist effect is included in the model to overcome this.

3.2.1 Multivariate extensions to the Bradley-Terry model

The two main non-parametric multivariate paired comparison tests for testing the hypothesis of no difference among several treatements are:

- the Sen-David test (Sen and David (1968))
- the Davidson-Bradley test (Davidson and Bradley (1969))

with a versatile and relatively simple approach to multivariate paired comparisons building on these methods being provided by Imray et al. (1976). A generalisation of Thurstonian probabilistic choice models for analysing both multiple preference responses and their relationships was proposed by Boeckenholt (1990).

Recent work by Dittrich et al. (2006) has developed a log-linear representation of the Bradley-Terry model for multivariate paired comparison data. By converting such data to multiple binomial responses, dependencies between the decisions of the judges, as well as possible association structures between the attributes, can be incorporated into the model, providing an advantage over parallel univariate analyses of individual attributes.

The Sen-David test

Suppose that the t objects $i = 1, \ldots, t$ are compared on the basis of two characteristics, say (ϵ, ζ). There are four possible outcomes for each set of comparisons:

$$k = 1: \quad \epsilon_i \rightarrow \epsilon_j; \zeta_i \rightarrow \zeta_j,$$

$$k = 2: \quad \epsilon_i \rightarrow \epsilon_j; \zeta_i \leftarrow \zeta_j,$$

$$k = 3: \quad \epsilon_i \leftarrow \epsilon_j; \zeta_i \rightarrow \zeta_j,$$

$$k = 4: \quad \epsilon_i \leftarrow \epsilon_j; \zeta_i \leftarrow \zeta_j,$$

where the direction of the arrow indicates preference. Thus in option 1 object i is chosen over object j on both attributes, whilst for option 2 i is chosen on ϵ and j is chosen on ζ.

Using these definitions, it is possible to define preference frequencies $N_{ij.k}\,(k = 1, \ldots, 4)$ which represents how often each comparison of objects i and j falls into each of the categories, along with the associated probabilities $p_{ij.k}$. Note that for all i, j and $k\,(i \neq j)$,

$$p_{ij.k} = p_{ji.5-k} \text{ and } N_{ij.k} = N_{ji.5-k}. \tag{3.3}$$

The overall N_{ij} comparisons (where N_{ij} is the total number of times objects i and j are compared), assumed independent, result in the multinomial distribution

$$\frac{N_{ij}!}{\prod_{k=1}^{4} N_{ij.k}!} \prod_{k=1}^{4} p_{ij.k}^{N_{ij.k}}. \tag{3.4}$$

The null hypothesis of equality of the objects with respect to both (ϵ, ζ) may be expressed as

$$p_{ij.k} = p_k, k = 1, \ldots, 4, \quad i \neq j = 1, \ldots, t,$$

which, by virtue of 3.3, means

$$p_1 + p_2 = p_1 + p_3 = p_2 + p_3 = p_3 + p_4 = \frac{1}{2}. \tag{3.5}$$

Equivalently, 3.5 may be written as

$$p_1 = p_4 = \frac{1}{4}(1 + \theta), \quad p_2 = p_3 = \frac{1}{4}(1 - \theta),$$

where θ is an association parameter $(-1 \leq \theta \leq +1)$. Correspondingly, the null hypothesis may therefore be written as

$$\begin{aligned} H_0 : p_{ij.k} &= \frac{1}{4}(1+\theta) \quad k = 1,4, \\ &= \frac{1}{4}(1-\theta) \quad k = 2,3, \text{ for all } i < j = 1,\ldots,t. \end{aligned} \tag{3.6}$$

Under H_0, the likelihood funcion of the entire sample is, from 3.4

$$\left(\prod_{i<j}^{t} \frac{N_{ij}!}{\prod_{k=1}^{4} N_{ij.k}!}\right) \frac{(1+\theta)^{N_1}(1-\theta)^{N-N_i}}{4^N}$$

where $N_1 = \sum_{i<j}^{t}(N_{ij.1} + N_{ij.4})$ and $N = \sum_{i<j}^{k} N_{ij}$.

Thus the maximum likelihood estimator of θ is simply

$$\hat{\theta}_N = \frac{2N_1 - N}{N} = \frac{N_1 - N_2}{N},$$

where $N_2 = N - N_1$, i.e. $\hat{\theta}_N$ is the difference in the proportions of like and unlike preferences for (ϵ, ζ).

In order to test H_O define

$$Z_{N,i}^{(1)} = \sum_{j}{}' N_{ij}^{-1}(N_{ij.1} - N_{ij.4})$$

and

$$Z_{N,i}^{(2)} = \sum_{j}{}' N_{ij}^{-1}(N_{ij.2} - N_{ij.3})$$

for $i \neq j = 1,\ldots,t$, and if any $N_{ij} = 0$, the corresponding term is omitted. Confining the discussion to a statement of the main results, let $\lim_{N\to\infty} \frac{N_{ij}}{N} = \rho_{ij}$ where $0 < \rho_{ij} < 1$ for all $i < j = 1,\ldots,t$.

Then for $|\theta| < 1$, under H_0 in 3.6, the test statistic

$$D_N = \frac{N}{t} \sum_{a=1}^{2} \frac{1}{N_a} \sum_{i=1}^{t} (Z_{N,i}^{a})^2$$

tends to a $\chi^2_{2(t-1)}$ distribution as $N \to \infty$. To facilitate extension from 2 to p attributes, define

$$T^{(1)}_{N,i} = Z^{(1)}_{N,i} + Z^{(2)}_{N,i}$$

and

$$T^{(2)}_{N,i} = Z^{(1)}_{N,i} - Z^{(2)}_{N,i}.$$

Then

$$D_N = \frac{1}{t\left(1-\hat{\theta}_N\right)} \sum_{i=1}^{t} \left[\left(T^1_{N,i}\right)^2 - 2\hat{\theta}_N T^1_{N,i} T^2_{N,i} + \left(T^2_{N,i}\right)^2 \right].$$

In the p-dimensional case $\frac{1}{2}p\,(p-1)$ association parameters corresponding to θ are needed. Define the $p \times p$ matrix $\Theta = \left(\hat{\theta}_{N.gh}\right)$, $g, h = 1, \ldots, p$, where $\hat{\theta}_{N.gh}$ is the estimate for attributes g and h, and $\hat{\theta}_{N.gg} = 1$, $g = 1, \ldots, p$. If $\hat{\theta}^{gh}_N$ is the $(g,h)^{th}$ element in Θ^{-1} then the test statistic is

$$D_{N(p)} = t^{-1} \sum_{g=1}^{p} \sum_{h=1}^{p} \hat{\theta}^{gh}_N \sum_{i=1}^{t} T^{(g)}_{N,i} T^{(h)}_{N,i},$$

which, under the null hypothesis of homogeneity of the objects, tends to a $\chi^2_{p(t-1)}$ distribution as $N \to \infty$.

The Davidson-Bradley test

Davidson and Bradley (1969) describe an extension of the univariate treatment parameters to p sets of treatment parameters, $\pi_{\alpha 1}, \ldots, \pi_{\alpha t}$, $\alpha = 1, \ldots, p$, $\pi_{\alpha i} \geq 0$, $i = 1, \ldots, t$, $\pi_{\alpha 1} + \ldots + \pi_{\alpha t} = 1$, and if $X_{\alpha i}$ and $X_{\alpha j}$ denote the paired response to treatments i and j on attribute α, then $p\,(X_{\alpha i} > X_{\alpha j}) = \frac{\pi_{\alpha i}}{\pi_{\alpha i} + \pi_{\alpha j}}$, $\alpha = 1, \ldots, p$, $i \neq j$, $i, j = 1, \ldots t$. As before, $X_{\alpha_i} > X_{\alpha j}$ is interpreted as choice of treatment i over treatment j, but now on attribute α.

Let $\mathbf{s} = (s_1, \ldots, s_p)$ be a preference or choice vector such that, when objects i and j are compared, $s_\alpha = i$ or j depending on whether object i or j is chosen on attribute α. Define the cell probability $p(\mathbf{s}|i,j)$ as the probability of the choice vector $\mathbf{s}$ for the (i,j) object pairing.

Now the multivariate model that has been selected may be formulated as follows: for each object pair (i,j) the cell probabilities are given by

$$p(\mathbf{s}|i,j) = p^{(1)}(\mathbf{s}|i,j)\, h(\mathbf{s}|i,j),$$

where

$$p^{(1)}(\mathbf{s}|i,j) = \prod_{\alpha=1}^{p} \frac{\pi_{\alpha s_\alpha}}{\pi_{\alpha i} + \pi_{\alpha j}},$$

$$h(\mathbf{s}|i,j) = 1 + \sum_{\alpha<\beta} \delta(s_\alpha, s_\beta)\, \rho\, (\pi_{\alpha i}/\pi_{\alpha j})^{\frac{1}{2}\delta(i,s_\alpha)} (\pi_{\beta i}/\pi_{\beta j})^{-\frac{1}{2}\delta(i,s_\beta)},$$

$s_\alpha = i, j$, $\alpha = 1, \ldots, p$ and $\delta(.,.) = \pm 1$, the sign being positive if the two arguments are equal and negative otherwise. The preference parameters, $\pi = \{\pi_{\alpha i}; i = 1, \ldots, t; \alpha = 1, \ldots, p\}$, are restrained by $\sum_{i=1}^{t} \pi_{\alpha i} = 1$, $\alpha = 1, \ldots, p$, and the parameters measuring association

$$\rho = \{\rho_{\alpha\beta}; \alpha < \beta, \alpha, \beta = 1, \ldots, p\},$$

are restricted by the requirement that $h(\mathbf{s}|i,j) \geq 0$ for each of the 2^p cells associated with each of the $\binom{t}{2}$ treatment comparisons. It is noted that $\rho = 0$ implies independence. Also it is interesting to note that, in the two-treatment, bivariate case, ρ_{12} is the ϕ-coefficient of correlation for the single 2x2 table.

Maximum-likelihood estimation is used to find the maximum-likelihood estimates of the model. The logarithm of the likelihood function is

$$\ell = \sum_{\alpha=1}^{p} \sum_{i=1}^{t} v_{\alpha i} \ln \pi_{\alpha i} - \sum_{\alpha=1}^{p} \sum_{i<j} N_{ij} \ln(\pi_{\alpha i} + \pi_{\alpha j}) + \sum_{i<j} \sum_{\mathbf{s}} f(\mathbf{s}|i,j) \ln h(\mathbf{s}|i,j)$$

where $f(\mathbf{s}|i,j)$ is the number of times the preference vector $\mathbf{s}$ occurs amongst the N_{ij} responses to the object pair (i,j), $v_{\alpha i} = \sum' n_{\alpha ij}$, the sum of the marginal frequencies of preference for object i over object j on attribute α, i.e. the number of times i is chosen over j on α, and $\sum_{\mathbf{s}}$ is the sum over the 2^p possible values of $\mathbf{s}$ representing the possible preference responses. The maximisation is subject to the constraints $\sum_{i=1}^{t} \pi_{\alpha i} = 1$ for all $\alpha = 1, \ldots, p$.

Davidson and Bradley (1971) considered an extension of the model looking at the problem of relating the response pattern on overall quality to that on a specified set of attributes. They derived a regression equation for a joint distribution of responses to dichotomous items, and applied it to the multivariate paired comparison model, along with a test of significance of the responses to specified attributes in estimating the responses to overall quality.

Example

Davidson and Bradley (1969) present an example based on a Chocolate pudding test, which is recreated here.

The responses to paired comparisons on attributes (1) taste or flavour, (2) colour, and (3) texture or feel in the mouth, for treatments B, C, and E, have been tabulated and the resulting set of cell frequencies, $f(\mathbf{s}|i,j)$ are presented in Table 3.2. These frequencies have been used to obtain the maximum likelihood estimates $\hat{\pi}$ of π and $\hat{\rho}$ of ρ. The estimated frequencies obtained by use of the maximum-likelihood estimates are also given in Table 3.2 in parentheses.

Treatment pair		Cell frequencies $f(\mathbf{s}\|i,j)$								Frequency
i	j	(iii)	(jii)	(iji)	(jji)	(iij)	(jij)	(ijj)	(jjj)	N_{ij}
1	2	8	1	1	1	0	2	0	9	22
		(7.93)	(1.09)	(1.15)	(1.69)	(0.76)	(0.97)	(0.37)	(8.03)	
1	3	6	0	1	1	1	0	1	9	19
		(6.25)	(0.60)	(1.24)	(0.92)	(1.12)	(0.62)	(0.64)	(7.61)	
2	3	7	1	1	1	3	1	1	6	21
		(6.92)	(0.37)	(1.26)	(0.60)	(1.70)	(0.75)	(1.10)	(8.31)	

Table 3.2: Observed and expected cell frequencies for chocolate pudding test, recreated from Davidson and Bradley (1969) Product B is labelled as 1, C as 2 and E as 3.

The set of final parameter estimates are presented in Table 3.3 and the correlation parameter estimates for the attributes are shown in Table 3.4. The likelihood ratio statistics have been computed for (*i*) the test of equal preferences on all three attributes in presence of correlation, (*ii*) the test of zero correlations, and (*iii*) the test of goodness of fit, following the methods given in Davidson and Bradley (1969). These are presented in Table 3.5 together with their significance test results.

α	$\hat{\pi_{\alpha B}}$	$\hat{\pi_{\alpha C}}$	$\hat{\pi_{\alpha E}}$
1	0.312	0.360	0.328
2	0.307	0.321	0.372
3	0.338	0.288	0.374

Table 3.3: Treatment parameter estimates for chocolate pudding test

$(\alpha\beta)$	$\hat{\rho}_{\alpha\beta}$
12	0.675
13	0.654
23	0.588

Table 3.4: Correlation parameter estimates for chocolate pudding test

Test	$-2\ln\lambda$	p-value
Equal preferrences	2.362	0.88
Zero correlations	62.665	< 0.0005
Goodness of fit	9.135	0.69

Table 3.5: Likelihood ratio statistics and tests of hypotheses based on chocolate pudding data

The results of this experiment show that there are no significant differences between the chocolate puddings on these three attributes (p=0.88). However, there is a high level of correlation between the attributes, as shown in Table 3.4.

In this Chapter a different approach is taken where an MDS type of analysis is used for visualising multivariate paired comparison data.

3.3 An MDS approach to multivariate paired comparison data

The aim of the technique is to produce a two-dimensional biplot-like representation showing how objects score on several attributes, as shown in the schematic in Figure 3.2. To produce such a plot requires calculation of both the coordinates of the points representing the objects, and the parameters of the axes for the attributes from the available data, which are the results of the various paired comparisons for each attribute.

For a bivariate paired comparison approach, the relationships between the attributes and between the objects can be represented simply by a two-dimensional scatter plot of the respective treatment parameters from the univariate Bradley-Terry test. However, increasing the number of attributes to $a > 2$ would mean that this approach would not be valid. Therefore, the focus of this Chapter is to detail a method for such instances, when $a > 2$.

Of course, representation of such multivariate data in only two dimensions will lead to some reduction in the information displayed, and a method to reduce this loss is sought. The process described here is used to set up a map with an arbitrary starting configuration. The treatment parameters can

then be determined for each attribute, and these used to calculate a pseudo-likelihood for the configuration. Iterative procedures will then allow this to be maximised, leading to the optimal configuration for the data. The maximisation function is not a true likelihood due to correlations between the attributes, and the possible effect of the joint distributions of the attributes.

Figure 3.2 shows a schematic of a map produced by the approach based on 5 objects and 3 attributes - each object being represented by a point, and each attribute by an axis.

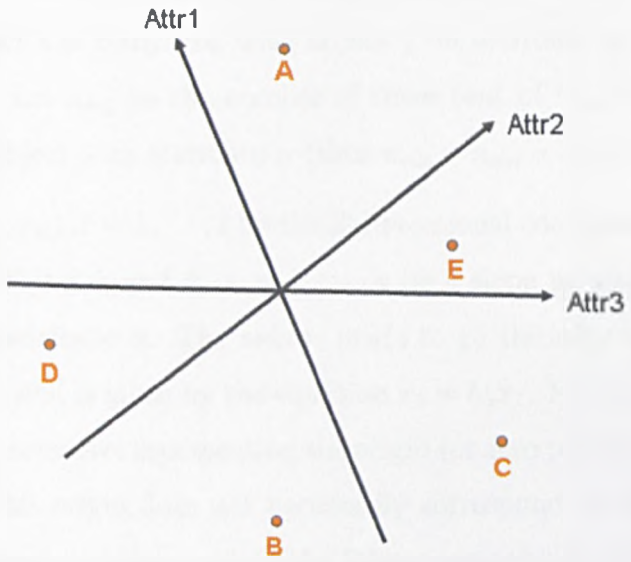

Figure 3.2: Schematic showing an MDS approach to visualising multivariate paired comparison data

The orthogonal projection from a point to the axis gives the relevant object rating parameter - i.e. $\pi_{\alpha i}$ is the object rating parameter for object i on attribute α. Thus the relative projection of two objects onto an axis shows the results of a paired comparison test on the two objects - the object projected furthest along the axis being chosen most often. For example in

Figure 3.2, when comparing objects A and C on attribute 1, object A has been chosen most often, whereas when comparing B and C on the same attribute, the objects are chosen equally often, as they project to the same point.

3.3.1 Calculation of the pseudo-likelihood

Assume that there are a attributes, and t treatments (or objects) compared in a pairwise manner on all of these attributes. Let $N_{\alpha ij}$ be the number of times object i is compared with object j on attribute α, $i, j = 1, \ldots, t$, $\alpha = 1, \ldots, a$. Let $n_{\alpha ij}$ be the number of times (out of $N_{\alpha ij}$) that object i is chosen over object j on attribute α (thus $n_{\alpha ij} + n_{\alpha ji} = N_{\alpha ij} = N_{\alpha ji}$).

Let $(x_{1i}, x_{2i}), i = 1, \ldots, t$ be the 2-dimensional coordinates of the point representing object i, and $b_\alpha, \alpha = 1, \ldots, a$ be a slope parameter of the axis representing attribute α. The axis is made to go through the origin of the configuration, and is given by the equation $x_2 = b_\alpha x_1$. Finally, define $c_\alpha, \alpha = 1, \ldots, a$ as a parameter representing the origin (or zero point) of the attribute on axis α. This origin does not necessarily correspond to the origin of the overall coordinate system, and so the former henceforth will be referred to as the zero point.

To calculate the pseudo-likelihood of this configuration, the first stage is to calculate the treatment parameters (i.e. $\pi_{\alpha i}$) for each attribute α. Figure 3.3 shows a schematic of how this is carried out for one attribute, α.

The object points (x_{1i}, x_{2i}) are projected onto the axis, at the projection point labelled $\lambda_{\alpha i}$. Now let the the relevant treatment parameter $\pi_{\alpha i}$ be defined as the distance along the axis from the zero point to the projection point $\lambda_{\alpha i}$.

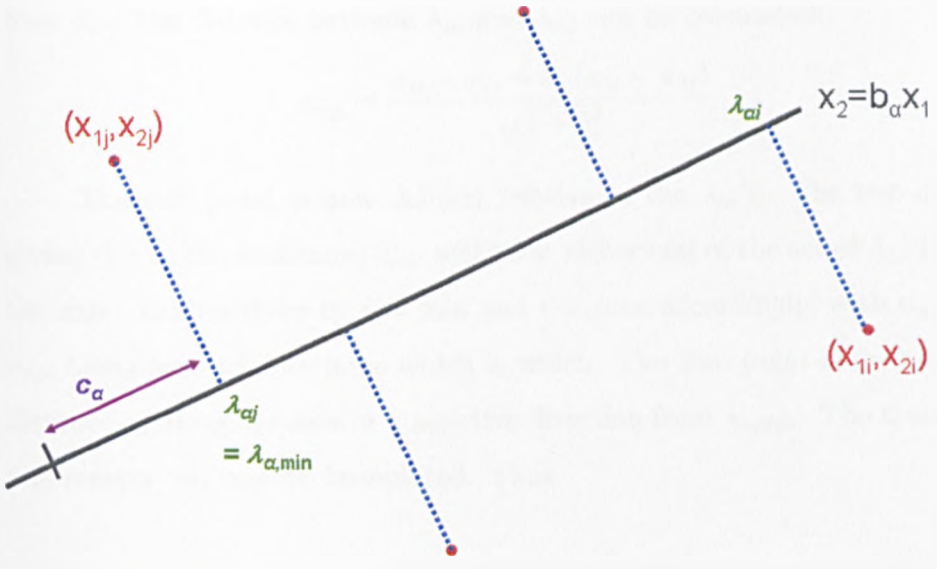

Figure 3.3: Schematic showing parameters needed for calculation of pseudo-likelihood

Let the coordinates of $\lambda_{\alpha i}$ be $\left(x_{1i}^{(p)}, x_{2i}^{(p)}\right)$, which can be rewritten as $\left(x_{1i}^{(p)}, b_\alpha x_{1i}^{(p)}\right)$. Now, by definition, $\lambda_{\alpha i}$ is located at the point where the distance between (x_{1i}, x_{2i}) and $\left(x_{1i}^{(p)}, b_\alpha x_{1i}^{(p)}\right)$ is a minimum. Let D be the distance between the points (x_{1i}, x_{2i}) and (x'_{1i}, x'_{2i}), an arbitrary point on the axis, and so

$$D^2 = (x_{1i} - x'_{1i})^2 + (x_{2i} - b_\alpha x'_{1i})^2 .$$

Differentiating and equating to zero gives

$$\frac{\mathrm{d}D^2}{\mathrm{d}x'_{1i}} = -2(x_{1i} - x'_{1i}) - 2b_\alpha (x_{2i} - b_\alpha x'_{1i}) = 0.$$

Solving this gives the coordinates $\left(x_{1i}^{(p)}, x_{2i}^{(p)}\right)$ as

$$x_{1i}^{(p)} = \frac{x_{1i} + b_\alpha x_{2i}}{1 + b_\alpha^2} \text{ and } x_{2i}^{(p)} = \frac{b_\alpha x_{1i} + b_\alpha^2 x_{2i}}{1 + b_\alpha^2}.$$

Now $d_{\alpha ij}$, the distance between $\lambda_{\alpha i}$ and $\lambda_{\alpha j}$ can be calculated:

$$d_{\alpha ij} = \frac{x_{1i} - x_{1j} + b_\alpha (x_{2i} - x_{2j})}{\sqrt{1 + b_\alpha^2}} \tag{3.7}$$

The zero point is now defined relative to the $\lambda_{\alpha i}$'s. The two objects giving rise to the maximum $d_{\alpha ij}$ will be at either end of the set of $\lambda_{\alpha i}$'s along the axis - denote these by $i = \min$ and $i = \max$ accordingly, with $n_{\alpha ij}$ and $n_{\alpha ji}$ being used to determine which is which. The zero point is defined as a distance c_α along the axis in a negative direction from $\lambda_{\alpha,\min}$. The treatment parameters can now be formulated. Thus

$$\pi_{\alpha i} = c_\alpha + d_{\alpha,\min,i} \quad i = 1, \ldots, t \tag{3.8}$$

where $d_{\alpha,\min,i}$ is the distance from $\lambda_{\alpha,\min}$ and λ_i. When $i = \min$, then $d_{\alpha,\min,i} = 0$. The treatment parameters are then scaled so that $\sum_i \pi_{\alpha i} = 1$ for each α.

The probability for chosing object i over object j on attribute α is

$$P_{\alpha ij} = \frac{\pi_{\alpha i}}{\pi_{\alpha i} + \pi_{\alpha j}}. \tag{3.9}$$

The pseudo-likelihood, L, for all attributes and treatments can therefore be written as

$$L = \prod_\alpha \prod_{i<j} \left(\frac{\pi_{\alpha i}}{\pi_{\alpha i} + \pi_{\alpha j}} \right)^{n_{\alpha ij}} \left(\frac{\pi_{\alpha j}}{\pi_{\alpha i} + \pi_{\alpha j}} \right)^{n_{\alpha ji}},$$

or

$$L = \prod_\alpha \prod_{i \neq j} \left(\frac{\pi_{\alpha i}}{\pi_{\alpha i} + \pi_{\alpha j}} \right)^{\frac{n_{\alpha ij}}{2}} \left(\frac{\pi_{\alpha j}}{\pi_{\alpha i} + \pi_{\alpha j}} \right)^{\frac{n_{\alpha ji}}{2}}.$$

Alternatively, the pseudo-likelihood can be written as

$$L = \prod_\alpha \frac{\prod_i (\pi_{\alpha i})^{\frac{n_{\alpha i.}}{2}}}{\prod_{i \neq j} (\pi_{\alpha i} + \pi_{\alpha j})^{\frac{N_{\alpha ij}}{2}}} \tag{3.10}$$

where $n_{\alpha i.} = \sum n_{\alpha ij}$, which is the total number of 'wins' for treatment i on attribute α.

Taking logarithms of equation 3.10 gives the pseudo-loglikelihood,

$$\ell = \ln L = \sum_{\alpha}\sum_{i} \frac{n_{\alpha i.}}{2} \ln(\pi_{\alpha i}) - \sum_{\alpha}\sum_{i \neq j} \frac{N_{\alpha ij}}{2} \ln(\pi_{\alpha i} + \pi_{\alpha j}). \qquad (3.11)$$

Combining Equations 3.8 and 3.11 gives

$$\ell = \sum_{\alpha}\sum_{i} \frac{n_{\alpha i.}}{2} \ln(c_\alpha + d_{\alpha,\min,i}) - \sum_{\alpha}\sum_{i \neq j} \frac{N_{\alpha ij}}{2} \ln(2c_\alpha + d_{\alpha,\min,i} + d_{\alpha,\min,j}).$$

and hence

$$\ell = \sum_{\alpha}\sum_{i} \frac{n_{\alpha i.}}{2} \ln\left[c_\alpha + \frac{x_{1i} - x^{(\alpha)}_{1,\min}}{\sqrt{1+b_\alpha^2}}\right]$$
$$- \sum_{\alpha}\sum_{i \neq j} \frac{N_{\alpha ij}}{2} \ln\left[2c_\alpha + \frac{x_{1i} + x_{1j} - 2x^{(\alpha)}_{1,\min} + b_\alpha\left(x_{2i} - x_{2j} - 2x^{(\alpha)}_{2,\min}\right)}{\sqrt{1+b_\alpha^2}}\right]$$

where $\left(x^{(p)}_{1,\min}, x^{(p)}_{2,\min}\right)$ are the coordinates of the point corresponding to $\lambda_{\alpha,\min}$.

Maximum-likelihood estimates of the coordinates of the points and the parameters of the axes can now be found by maximising ℓ. This was done iteratively using the double dogleg optimisation method found in PROC IML in the SAS software package. This optimisation method combines the ideas of the quasi-Newton and trust-region methods (Dennis and Mei (1979), Gay (1983) and Fletcher (1987)).

An example

Here, an example is given based on the chocolate pudding data from Davidson and Bradley (1969), as seen in Section 3.2.1. Initially, a univariate approach for each attribute in turn gives the results shown in Table 3.6.

Attribute 1: Taste		Attribute 2: Colour		Attribute 3: Texture	
Product	π_i	Product	π_i	Product	π_i
C	0.409	C	0.388	E	0.379
E	0.309	E	0.354	C	0.320
B	0.281	B	0.261	B	0.299
p-value = 0.5495		p-value = 0.5350		p-value = 0.8031	

Table 3.6: Univariate Bradley-Terry results for chocolate pudding data

This confirms the earlier results in that there are no significant differences between the products. There also appears to be correlations between the attributes - note especially the order of the products for attributes 1 and 2.

Applying the MDS-approach to the data results in the plot shown in Figure 3.4. This configuration has a pseudo-loglikelihood of $\ell = -127.474$. Looking at the projection of the points onto the axes, it can be seen that the univariate order is maintained, i.e. Attribute 1 has $C > E > B$, Attribute 2 also has $C > E > B$, whilst Attribute 3 has $E > C > B$.

The treatment parameters calculated by the MDS-approach are shown in Table 3.7. They are identical to the univariate treatment parameters shown in Table 3.6. The reason why this occurs is because to fit three products to one axis, only two parameters are actually involved, and this means a perfect fit can be achieved in two dimensions (with the perfect fit here being defined as the treatment parameters derived from the univariate Bradley-Terry test). The necessary parameters are the zero point (which defines, say $\pi_{\alpha 1}$) and the ratio between the distances, say $d_{\alpha 12}$ and $d_{\alpha 13}$, which allows definition of $\pi_{\alpha 2}$ and $\pi_{\alpha 3}$. Thus there are six parameters and six coordinates.

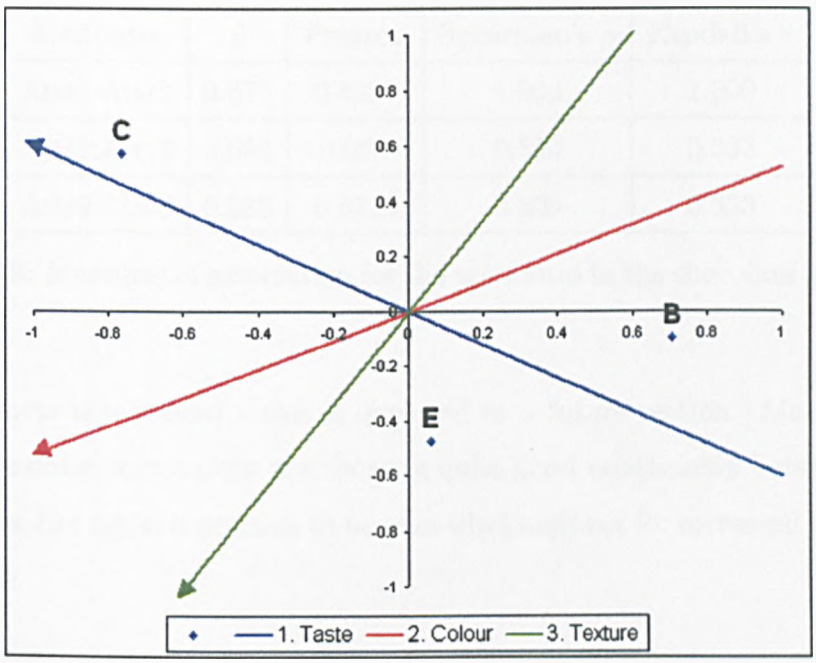

Figure 3.4: MDS-approach for displaying chocolate pudding data

Attribute	π_B	π_C	π_E
1	0.281	0.409	0.309
2	0.261	0.383	0.354
3	0.299	0.320	0.379

Table 3.7: Treatment parameters calculated by MDS-approach method

Table 3.8 shows a number of correlation measures between the attributes. Also included are the $\hat{\rho}$ measures of association from the Davidson and Bradley method.

Whilst it appears the Pearson correlation is not giving the same results as the $\hat{\rho}$, it should be remembered that this correlation is based on three points, so is not very robust. It is expected that the Pearson correlation measure will give a better measure of association when the number

Attributes	$\hat{\rho}$	Pearson	Spearman's ρ	Kendall's τ
Attr1:Attr2	0.675	0.8257	1.000	1.000
Attr1:Attr3	0.654	-0.0504	0.500	0.333
Attr2:Attr3	0.588	0.5128	0.500	0.333

Table 3.8: Measures of association for the attributes in the chocolate pudding test

of products is increased - this is explored in a future section. Meanwhile, the rank-order correlations are showing quite good relationship between the products, but again it remains to be seen what happens for increased product numbers.

3.3.2 Extending the methodology

The methodology is not limited to linear axes, nor a 2-dimensional representation. Here, a generalisation of the method is presented.

The treatment points have coordinates $(x_{1i}, \ldots, x_{mi})$ in an m- dimensional map, and are projected onto the axis, at the projection points labelled $\lambda_{\alpha i}$. The axis is now defined as a regular curve $(f_{1\alpha}(t), f_{2\alpha}(t), \ldots, f_{m\alpha}(t))$.

Let the object closest to the zero point be denoted by $i = \min$ with points $\lambda_{\alpha,\min}$ and $\pi_{\alpha,\min}$. Again the other treatment parameters $\pi_{\alpha i}$ are defined as the distance along the axis from the zero point to the projection point $\lambda_{\alpha i}$.

In Figure 3.5, object r projects closest to the zero point, and so this object has projection point $\lambda_{\alpha,\min}$ and treatment parameter $\pi_{\alpha,\min}$. This schematic represents a two-dimensional solution, but the principle is the same for $m > 2$ dimensions. The subsequent treatment parameters are then de-

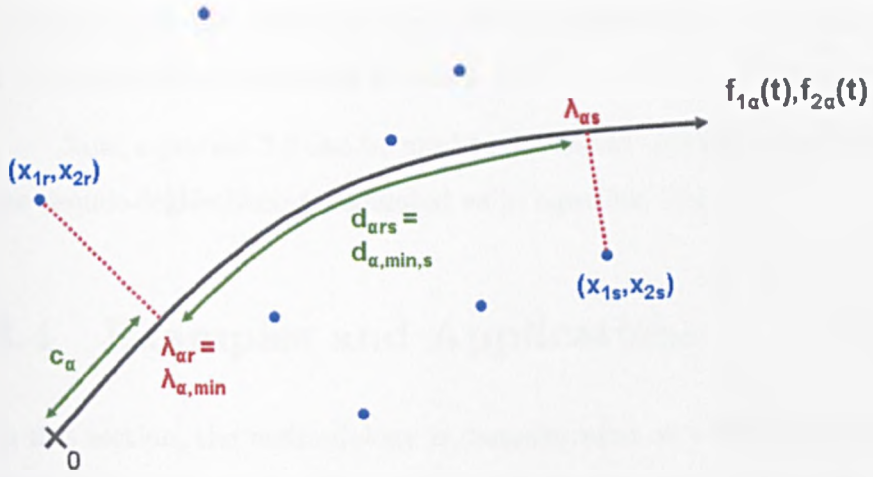

Figure 3.5: Schematic showing parameters needed for calulcation of Pseudo-likelihood

termined by the distance between its projection point and the zero point, i.e.

$$\pi_{\alpha i} = c_\alpha + d_{\alpha,\min,i} \quad i = 1, \ldots, t,$$

Let the coordinates of $\lambda_{\alpha i}$ be $\left(x_{1i}^{(p)}, \ldots, x_{mi}^{(p)}\right)$, and $t_{\alpha i}$ be the value of t at $\lambda_{\alpha i}$. Now following the same argument given in Section 2.2.1, the equation

$$\sum_{k=1}^{m} f'_{k\alpha}(t_{\alpha i}) \left[f_{k\alpha}(t_{\alpha i}) - x_{ki}\right] = 0 \tag{3.12}$$

can be solved to find the value of $t_{\alpha i}$ for all $i = 1, \ldots, t$.

The next stage is to calculate the lengths between all the possible pairs of the projections points, i.e.

$$d_{\alpha ij} = \left| \int_{t_{\alpha i}}^{t_{\alpha j}} \sqrt{\sum_{k=1}^{m} f_{k\alpha}^{t}(t)^2} \, \mathrm{dt} \right| \tag{3.13}$$

Once the $d_{\alpha ij}$ have been calculated, the next stage is to determine which $\lambda_{\alpha i}$ is $\lambda_{\alpha\,\min}$. This is carried out by looking for $d^* = \max(d_{\alpha ij})$, which

corresponds to the two projection points furthest apart, and using $n_{\alpha ij}$ and $n_{\alpha ji}$ to determine which end is which.

Now, equation 3.8 can be used to define the treatment parameters, and the pseudo-loglikelihood calculated as in equation 3.10.

3.4 Examples and Applications

In this section, the methodology is demonstrated on a simulated data set to show how well it works, and then also on real data sets. Whilst the technique is capable of handling a large number of objects, here the examples are limited to just five objects. This is because the methodology is designed for analysing sensory data, and when carrying out paired comparisons in sensory testing, 5 objects is often said to be the limit. Even with 5 objects, each panellist is being asked to make $\binom{5}{2} = 10$ individual comparisons. Any increase can lead to panellist fatigue.

3.4.1 Simulated data

Simulation of the data

A 2-dimensional map was constructed with 5 objects, with coordinates in both dimensions randomly selected from a uniform distribution between -0.5 and + 0.5. Three linear axes were then selected to represent three attributes. The axes were parameterised as $f_{1\alpha}(t) = (b_{\alpha 1}t, b_{\alpha 2}t)$ with the values for $b_{\alpha 1}$ and $b_{\alpha 2}$ being selected randomly from a uniform distribution between -5 and +5. The zero values, c_α were all set to be 0.001. This forms a 'perfect model', which the methodology should be able to recreate perfectly. The addition of noise to the perfect model is dealt with later.

The simulated plot can be seen in Figure 3.6. The axes are given by: $(f_{11}(t) = 0.901t, f_{21}(t) = 0.115t)$, $(f_{12}(t) = -3.232t, f_{22}(t) = 2.431t)$, and $(f_{13}(t) = 0.756t, f_{23}(t) = 1.043t)$, whilst the coordinates for the objects A, B, C, D and E are respectively $(0.256, 0.264)$, $(-0.231, -0.232)$, $(-0.142, 0.365)$, $(0.396, -0.432)$ and $(-0.279, 0.035)$.

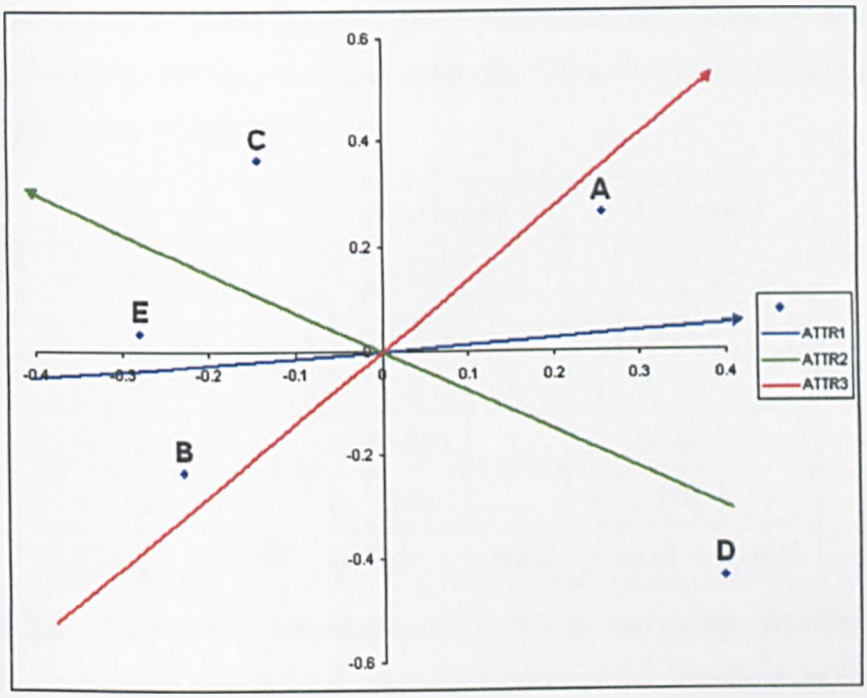

Figure 3.6: Simulated map displaying multivariate paired comparisons

Using equation 3.12, the projection points, $\lambda_{\alpha i}$, for this configuration can be calculated, followed by the distances between each pair of projection points, $d_{\alpha ij}$, from equation 3.13. The treatment parameters, $\pi_{\alpha i}$ are then calculated using equation 3.8.

As the treatment parameters are known, the probability of choosing treatment i over treatment j is also known, from equation 3.9. If in the simulated data, it is assumed that 50 subjects made the original comparisons

(i.e. $N_{\alpha ij} = 50$ for all α,i and j), then some data are simulated by

$$n_{\alpha ij} = 50\frac{\pi_{\alpha i}}{\pi_{\alpha i} + \pi_{\alpha j}}.$$

Calculations based on simulated data

Carrying out a univariate Bradley-Terry analysis on each simulated attribute in turn results in the treatment parameters in Table 3.9, with the parameters being plotted in Figure 3.7.

Attribute 1			Attribute 2			Attribute 3		
D	0.383	a	C	0.296	a	A	0.375	a
A	0.352	a	E	0.274	a	C	0.296	a
C	0.149	b	B	0.216	ab	D	0.146	b
B	0.062	c	A	0.183	b	E	0.135	b
E	0.054	c	D	0.031	c	B	0.048	c
p-value < 0.0001			p-value < 0.0001			p-value < 0.0001		

Table 3.9: Treatment parameters calculated by univariate Bradley-Terry model. Lower case letters join treatments that are not significantly different on that attribute at the 95% level of confidence

As can be seen in Figure 3.7, for attribute 1 object D scores the highest, followed by A, then C, with B and E being the lowest scoring objects. Thus it is expected that this order is maintained on the multivariate plot.

Figure 3.8 shows the plot generated by the methodology. As can be seen the plot has been rotated 90° clockwise from that produced from the generating data. This is because the orientation of the map produced is not unique. In addition, the locations of the points have moved slightly, but

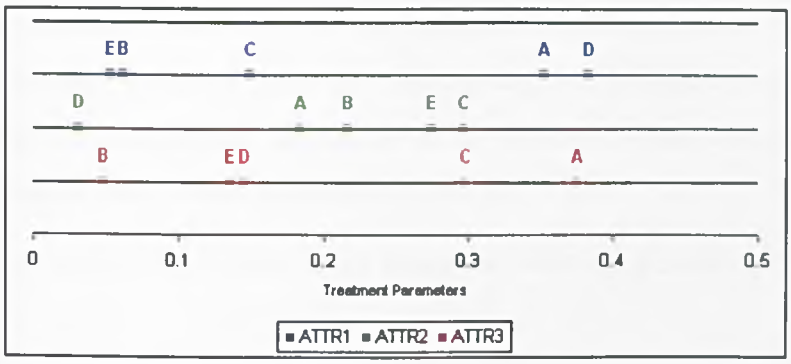

Figure 3.7: Univariate treatment parameters calculated from simulated data

the order of their projections has stayed exactly the same. For instance for attribute 1, the objects project in the order D, A, C, B and then E.

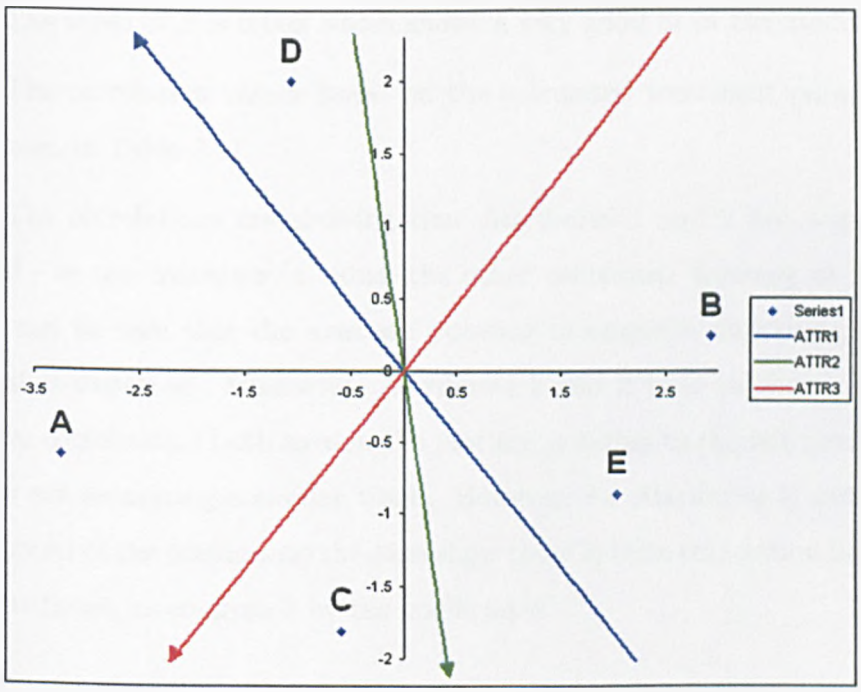

Figure 3.8: Fitted map for simulated data

The pseudo-loglikelihood for this configuration is -825.751. One way

to measure the goodness-of-fit of this configuration is to compare the calculated treatment parameters from the methodology with the known treatment parameters produced in the calculation of the method, which match the univariate Bradley-Terry scores given in Table 3.9.

These scores can be compared using the following equation:

$$S = \frac{\sum_{\alpha}^{a} \sum_{i}^{t} |\hat{\pi}_{\alpha i} - \pi_{\alpha i}|}{at}. \tag{3.14}$$

where $\hat{\pi}_{\alpha i}$ is the calculated the treatment parameter, and $\pi_{\alpha i}$ is the true treatment parameter. The lower the value of S, the better the configuration represents the data. The values of $\hat{\pi}_{\alpha i}$ and $\pi_{\alpha i}$ are shown in Table 3.10.

The value of S is 0.002 which shows a very good fit of the model.

The correlation values based on the calculated treatment parameters are shown in Table 3.11.

The correlations are showing that Attributes 1 and 2 are negatively related - as one increases in value, the other decreases. Looking at Figure 3.8 it can be seen that the axes are pointing in opposite directions, which is what is expected. Meanwhile Attributes 1 and 3 tend to show a slight positive correlation - both axes in the plot are pointing to the left hand side, and so are measuring a similar trend. However, for Attributes 2 and 3 the projections of the points onto the axes show there is little correlation between the attributes, as confirmed by the coefficients.

α	i	$\hat{\pi}_{\alpha i}$	$\pi_{\alpha i}$
1	1	0.354	0.352
1	2	0.061	0.062
1	3	0.149	0.149
1	4	0.381	0.383
1	5	0.053	0.054
2	1	0.186	0.183
2	2	0.213	0.216
2	3	0.298	0.296
2	4	0.029	0.031
2	5	0.272	0.274
3	1	0.371	0.375
3	2	0.047	0.048
3	3	0.300	0.296
3	4	0.144	0.146
3	5	0.136	0.135

Table 3.10: Calculated and 'real' treatment parameters from simulated data

Comparison	Pearson	Spearman	Kendall
Attr1:Attr2	-0.7579	-0.7000	-0.6000
Attr1:Attr3	0.4997	0.6000	0.4000
Attr2:Attr3	0.1864	-0.1000	0.0000

Table 3.11: Correlation results on attributes from simulated data

3.4.2 Visualising a multivariate paired comparison sensory test on deodorants

A sensory test looking at the characteristics of several deodorants is used to illustrate the methodology. Fifteen assessors participated in this sensory experiment aiming at an analysis of the comparison of deodorants. Five different deodorants were used for this experiment. The deodorants were applied to each panellist in pairs (randomised across the two armpits), and the panellist asked to compare them on several sensory attributes.

To demonstrate the MDS approach to visualising such data, these data will be used to generate a map.

Table 3.12 shows the treatment parameters generated by the univariate Bradley-Terry model, with the values plotted in Figure 3.9.

Attribute	A	B	C	D	E	p-value
Attribute 1	0.252 ab	0.459 a	0.035 c	0.122 b	0.132 b	<0.0001
Attribute 2	0.216 a	0.188 a	0.196 a	0.250 a	0.151 a	0.7543
Attribute 3	0.103 a	0.216 a	0.298 a	0.178 a	0.205 a	0.0668
Attribute 4	0.307 ab	0.374 a	0.048 d	0.114 c	0.157 bc	<0.0001
Attribute 5	0.168 a	0.124 a	0.263 a	0.207 a	0.237 a	0.2418
Attribute 6	0.088 b	0.189 a	0.293 a	0.243 a	0.187 ab	0.0162
Attribute 7	0.372 a	0.204 a	0.052 c	0.160 b	0.212 ab	<0.0001
Attribute 8	0.189 a	0.123 a	0.232 a	0.182 a	0.275 a	0.2618
Attribute 9	0.095 b	0.177 ab	0.313 ab	0.228 ab	0.187 ab	0.0237

Table 3.12: Treatment parameters calculated by univariate Bradley-Terry model on deodorant sensory data. Lower case letters join treatments that are not significantly different on that attribute at the 95% level of confidence.

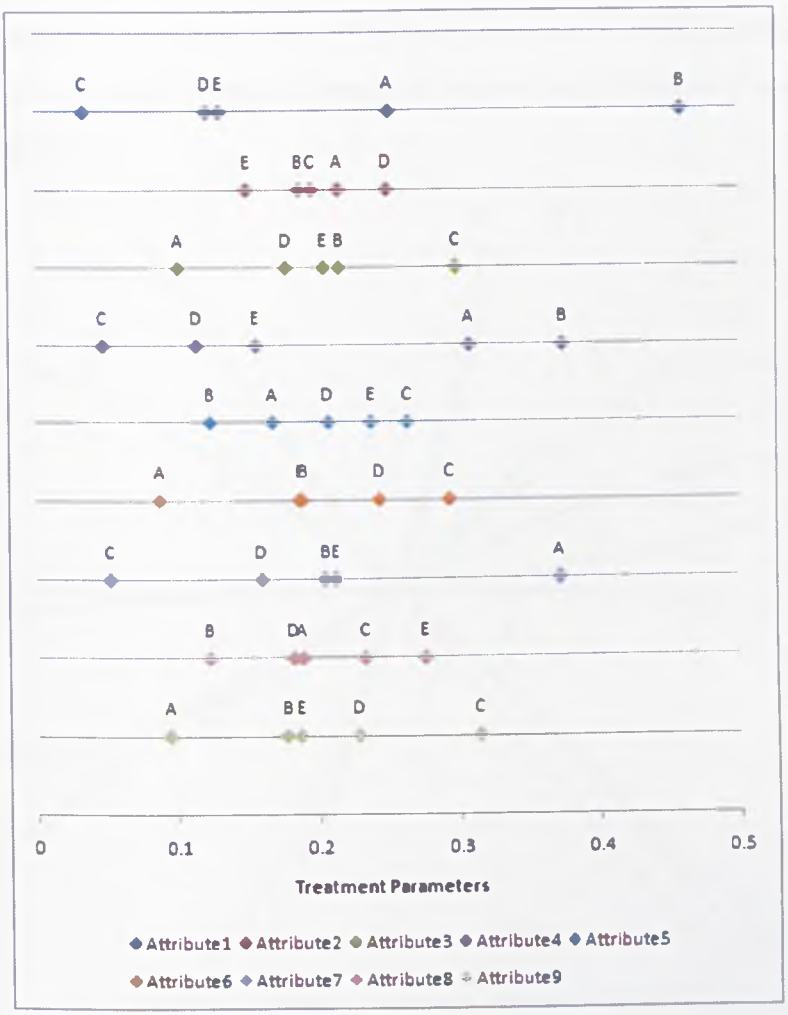

Figure 3.9: Plots of univariate treatment parameters for deodorant sensory data

Looking at the univariate results, it can be seen that there are no significant differences between treatments for the attributes 2, 5 and 8. Therefore these will be difficult for the multivariate methodology to fit. Deodorant C is always scored highest on attributes 3, 6 and 9, with A being the lowest. Deodorants B,D, and E are in the middle of this scale. Therefore, the three axes representing the these three attributes should be close together, due to

the high correlation of the scores. Finally, deodorant C is always the lowest on attributes 1, 4 and 7, with B being the highest on attributes 1 and 4, and A on attribute 9.

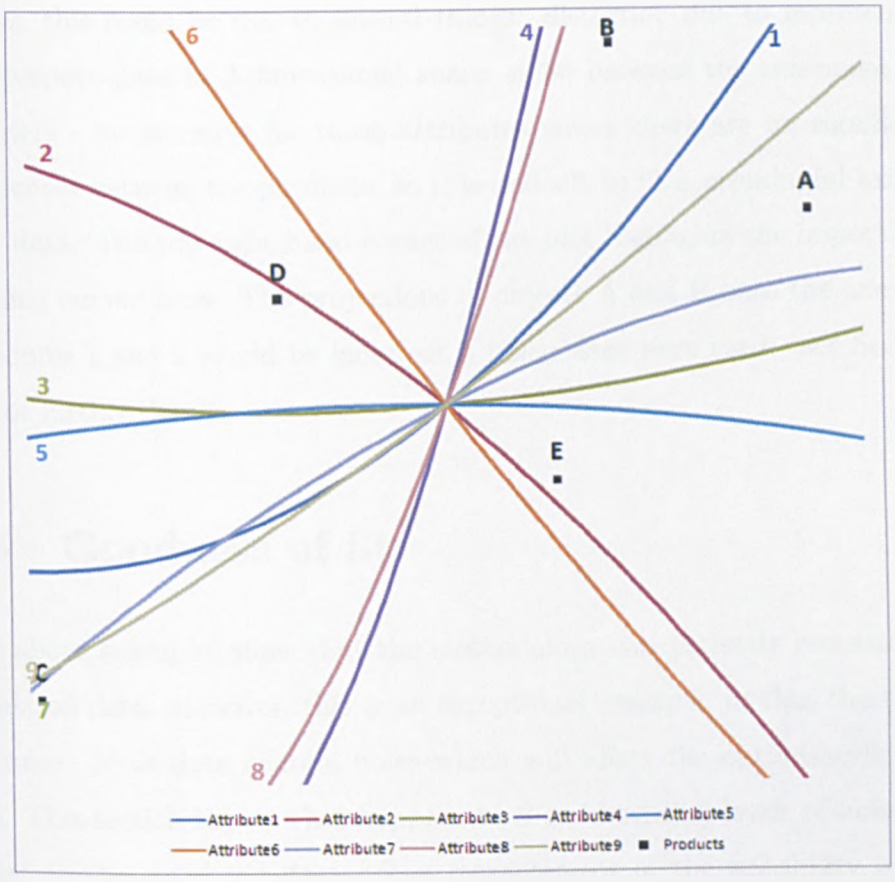

Figure 3.10: Results of MDS approach for plotting deodorant sensory multivariate paired comparison data. Numbered labels on axes are at the higher end of the scales

The plot resulting from fitting quadratic axes by the multivariate analysis can be seen in Figure 3.10. The pseudo-loglikelihood value for this configuration is $\ell = -771.117$. Study of this plot reveals that the projected values are as expected. For example, the axis for attribute 1 has product

C projecting at the lowest point, followed by D, E and A, with B being the highest projected point. This matches the order found in Table 3.12. Where there is less agreement between the plot and the univariate Bradley-Terry values, this could be due to several things: distortion due to representing multivariate data in 2-dimensional space; noise between the treatment parameters - for example for those attributes where there are no significant differences between the products, so it is difficult to fit a meaningful axis to such data. The top right hand corner of the plot highlights the importance of using curved axes. The projections of objects A and B onto the axes for attributes 1 and 9 would be incorrect if linear axes were used - see Section 3.6 for further details.

3.5 Goodness of fit

The above examples show that the methodology can perfectly recreate the simulated data. However, this is an exceptional example, in that there was no noise. Most data contain noise which will affect the optimisation process. This section shows what happens to S as increasing levels of noise are added to the simulated data. S, a measurement of the suitability of the configuration, was introducted in Section 3.4,

$$S = \frac{\sum_{\alpha}^{a} \sum_{i}^{t} |\hat{\pi}_{\alpha i} - \pi_{\alpha i}|}{at}.$$

The smaller the value of S, the better the configuration.

Equation 3.15 shows how noise was added to the model.

$$n_{\alpha ij} \sim Bn\left(N_{\alpha ij}, \frac{\pi_{\alpha i}}{\pi_{\alpha i} + \pi_{\alpha j}}\right), \tag{3.15}$$

for $\alpha = 1, \ldots, a$, $i = 2, \ldots, t$, and $j = 1, \ldots, i$. In other words, the $n_{\alpha ij}$ were randomly chosen from a binomial distribution with the probability

being calculated from the treatment parameters. $N_{\alpha ij}$ was set to be 50, and balance within the raw data maintained by $n_{\alpha ji} = N_{\alpha ij} - n_{\alpha ij}$.

One thousand random samples were thus generated. The distributions of the pseudo-loglikelihood and S are shown in Figure 3.11 with summary statistics in Table 3.13.

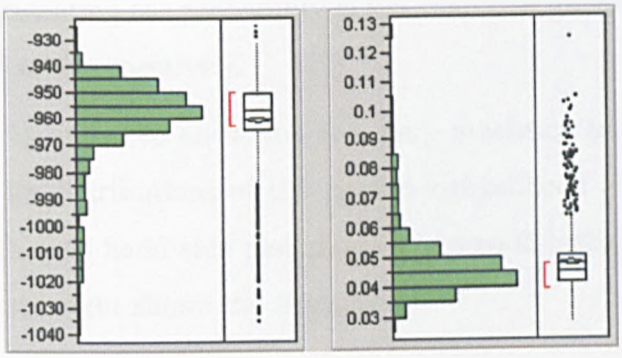

Figure 3.11: Distribution of fit statistics from methodology when noise is added by a binomial process. The pseudo-loglikelihood values are shown on the left hand side, and S on the right hand side

PLL mean	PLL std dev	PLL max	PLL min	S mean	S std dev	S max	S min
-960.6	16.77	-925	-1035	0.048	0.0119	0.1254	0.0288

Table 3.13: Summary of fit statistics from methodology when noise is added by a binomial process

The mean value for S is 0.048, and the distribution shows a positive skew. This is what would be expected from a well fitting configuration (compare though the value of S of 0.02 when there was no noise at all to see the effect adding some noise has had).

To show increasing levels of noise, the following was used for a random

sample of r of the 30 $n'_{\alpha ij}s$:

$$n_{\alpha ij} \sim Bn\left(N_{\alpha ij}, \frac{\pi_{\alpha i}}{\pi_{\alpha i} + \pi_{\alpha j}} + \kappa\right),$$

where $\kappa \sim U(-k, +k)$. To show different levels of noise, the following values were chosen: r=1,2,5,10 and 20, and k=0.0001, 0.001, 0.01, 0.1, 0.2, and 0.5. If the term representing the probability is less than 0 or greater than 1, then it is set to be 0 or 1 respectively.

Tables 3.14 and 3.15 show the summary statistics and Figures 3.12 to 3.16 show the distributions of the pseudo-loglikelihood and S from the simulations. The left hand side plot shows the pseudo-loglikelihood values, whilst that on the right shows the S value.

k	r	Mean	Std.Dev	Max	Min		r	Mean	Std.Dev	Max	Min
0.0001	1	-960.08	15.602	-929	-1040		10	-960.85	17.058	-927	-1040
0.001	1	-959.96	15.771	-928	-1034		10	-960.78	15.784	-930	-1027
0.01	1	-961.03	17.715	-918	-1048		10	-960.64	16.555	-930	-1033
0.1	1	-960.92	16.471	-928	-1036		10	-959.35	15.996	-927	-1042
0.2	1	-960.84	16.930	-925	-1045		10	-958.22	18.945	-913	-1037
0.5	1	-960.59	17.265	-925	-1046		10	-957.43	25.047	-889	-1070
0.0001	2	-960.69	17.148	-926	-1044		20	-959.75	16.017	-928	-1058
0.001	2	-960.33	16.029	-928	-1033		20	-959.99	15.743	-923	-1035
0.01	2	-960.90	17.132	-924	-1035		20	-960.06	15.808	-930	-1042
0.1	2	-960.35	16.916	-928	-1038		20	-958.89	18.200	-910	-1039
0.2	2	-960.74	18.125	-927	-1038		20	-956.97	21.336	-899	-1037
0.5	2	-958.80	17.270	-926	-1042		20	-953.24	32.543	-836	-1043
0.0001	5	-960.12	15.848	-920	-1044						
0.001	5	-959.85	15.900	-929	-1037						
0.01	5	-960.60	16.221	-929	-1035						
0.1	5	-959.87	17.487	-927	-1038						
0.2	5	-959.80	18.603	-907	-1041						
0.5	5	-959.64	20.608	-902	-1034						

Table 3.14: Pseudo-loglikelihood parameters from adding noise to simulated data set

k	r	Mean	Std.Dev	Max	Min	r	Mean	Std.Dev	Max	Min
0.0001	1	0.0474	0.01146	0.117	0.030	10	0.0479	0.01261	0.108	0.027
0.001	1	0.0474	0.01156	0.109	0.027	10	0.0479	0.01168	0.100	0.027
0.01	1	0.0483	0.01286	0.106	0.030	10	0.0479	0.01194	0.118	0.028
0.1	1	0.0484	0.01243	0.108	0.026	10	0.0483	0.01153	0.105	0.027
0.2	1	0.0486	0.01234	0.107	0.029	10	0.0522	0.01174	0.108	0.029
0.5	1	0.0497	0.01220	0.118	0.024	10	0.0666	0.01599	0.141	0.028
0.0001	2	0.0480	0.01245	0.115	0.028	20	0.0471	0.01145	0.115	0.028
0.001	2	0.0476	0.01153	0.111	0.030	20	0.0473	0.01133	0.116	0.029
0.01	2	0.0477	0.01218	0.109	0.025	20	0.0471	0.01127	0.109	0.028
0.1	2	0.0480	0.01232	0.112	0.028	20	0.0499	0.01177	0.101	0.029
0.2	2	0.0490	0.01254	0.108	0.031	20	0.0556	0.01260	0.108	0.026
0.5	2	0.0509	0.01218	0.113	0.028	20	0.0804	0.01846	0.146	0.026
0.0001	5	0.0469	0.01195	0.116	0.025					
0.001	5	0.0474	0.01134	0.106	0.030					
0.01	5	0.0477	0.01219	0.109	0.030					
0.1	5	0.0484	0.01245	0.110	0.025					
0.2	5	0.0502	0.01218	0.103	0.027					
0.5	5	0.0573	0.01369	0.111	0.027					

Table 3.15: S (fitting) parameters from adding noise to simulated data set

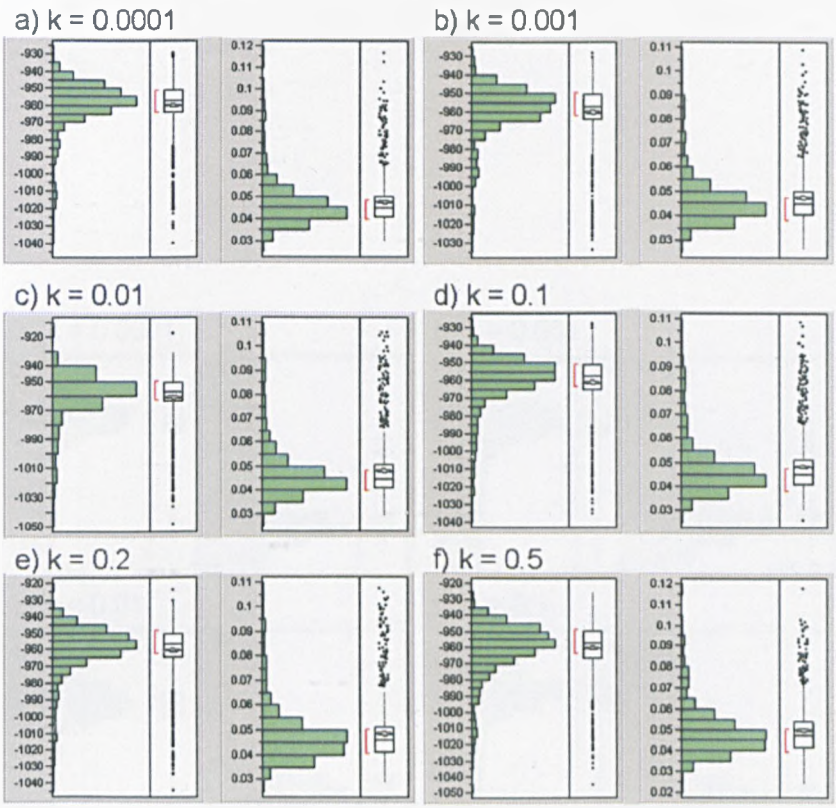

Figure 3.12: Effect of adding noise when $r = 1$

The results show that overall the methodology is very robust. When the level of noise, or the number of affected points is low, then the S statistic is still very close to the no noise value, with very little spread in the data. In fact, it is not until $r = 5$ and $k = 0.5$ that the method has trouble finding a configuration.

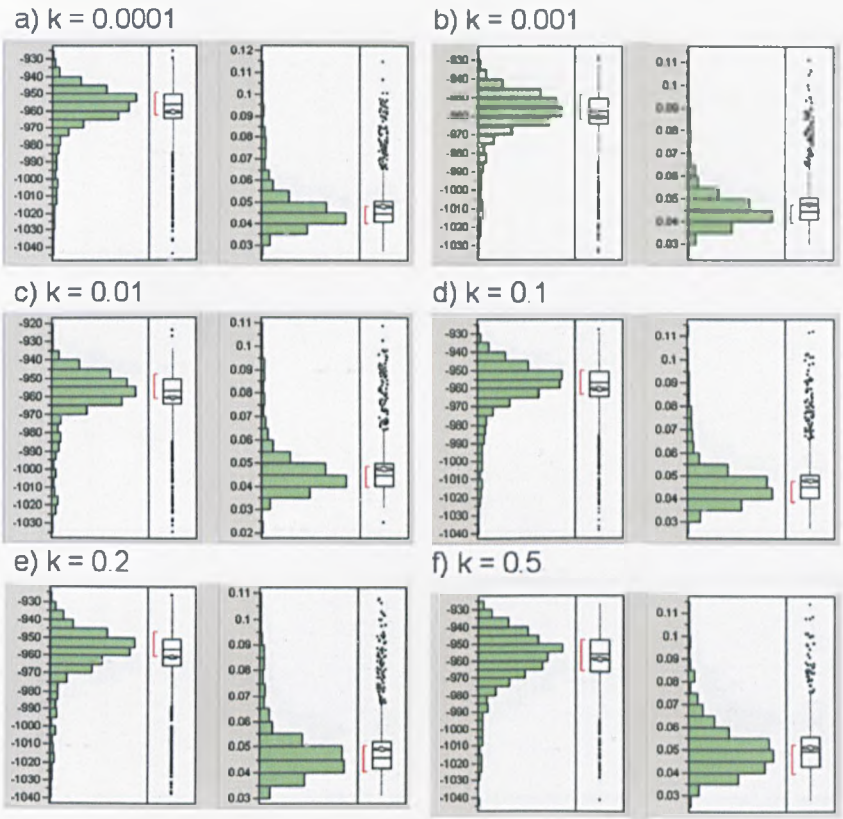

Figure 3.13: Effect of adding noise when $r = 2$

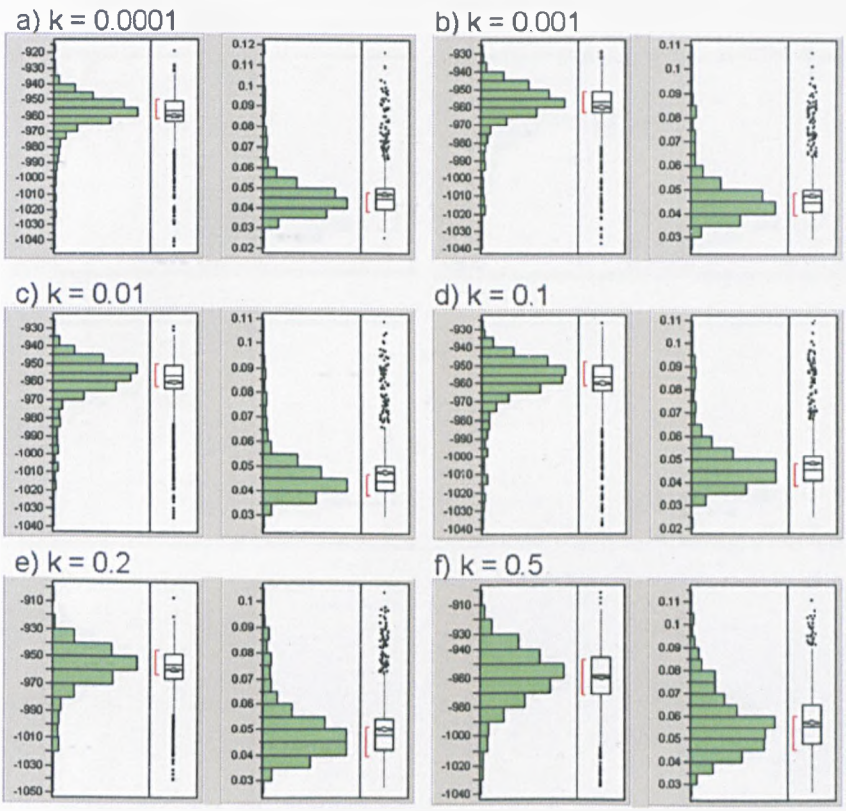

Figure 3.14: Effect of adding noise when $r = 5$

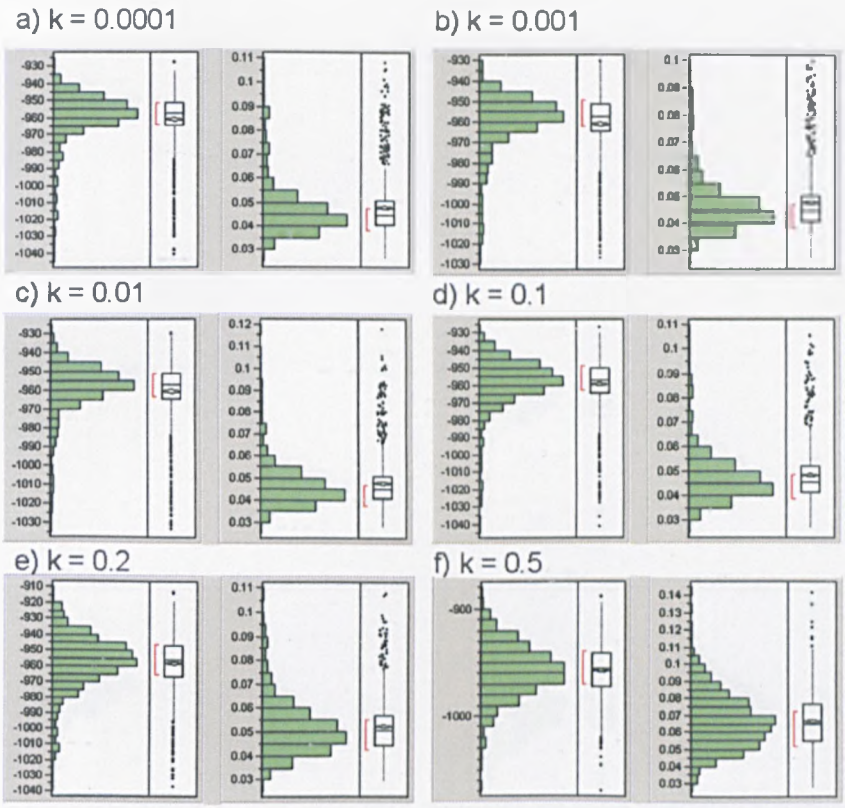

Figure 3.15: Effect of adding noise when $r = 10$

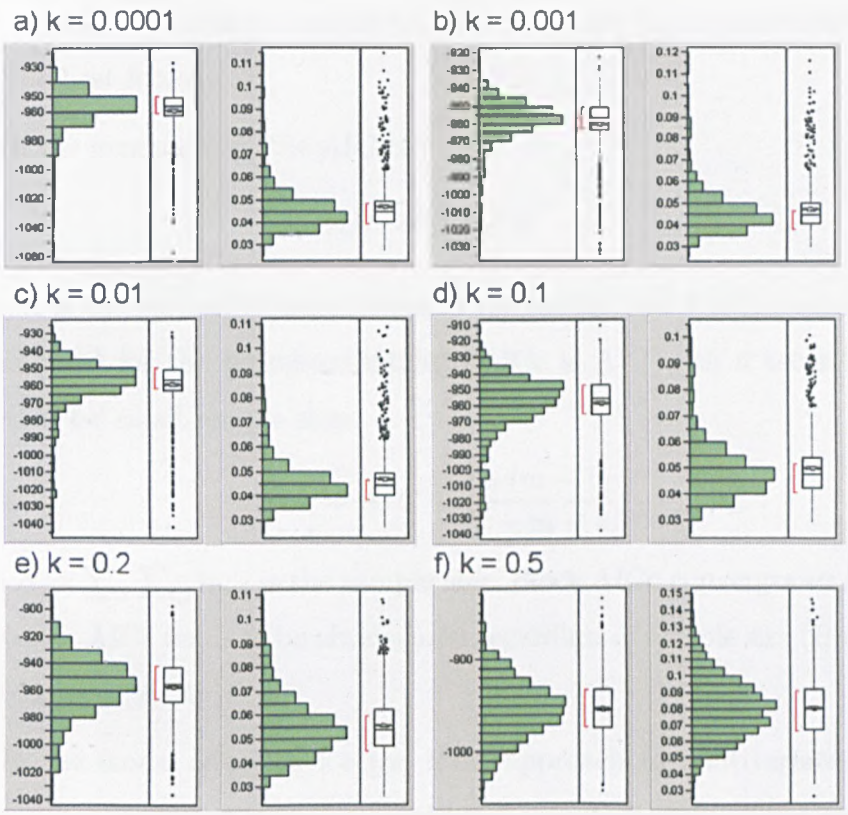

Figure 3.16: Effect of adding noise when $r = 20$

3.6 Model selection

As the configuration is optimised using the (pseudo) maximum likelihood function, it seems obvious to use this in the selection of the functions to describe the axes. This instinctively leads to an Akaike Information Criterion (AIC) approach (Akaike (1974)). Given a data set, several competing models may be ranked according to their AIC, with the one having the lowest AIC being the best fitting.

In the general case, the AIC is

$$AIC = 2m - 2\ell$$

where m is the number of parameters in the model, and ℓ is the maximised log-likelihood for the estimated model. AICc is AIC with a second order correction for small sample sizes,

$$AICc = AIC + \frac{2m(m-1)}{N-m-1},$$

where $N = \sum_{\alpha} \sum_{i,j} n_{\alpha ij}$ is the sample size. Since AICc converges to AIC as n increases, AICc tends to be always used regardless of sample size (Burnham and Anderson (2004b)).

In the model selection for the MDS approach to multivariate paired comparisons, all of the axes are considered simultaneously. Thus m is the total number of parameters in all of the axes. The sample size n is chosen to be the total number of paired comparisons $a\binom{t}{2}$, where a is the number of attributes, and t the number of treatments.

Taking the simulated data from Section 3.4, the following axes were fitted:

- All axes linear: $f_{1i}(t) = b_{1i}t$ and $f_{2i}(t) = b_{2i}t$ for all i.

- Axis 1 quadratic; Axes 2 and 3 linear: $f_{11} = b_{11}t$ and $f_{21} = b_{21}t + b_{31}t^2$; and $f_{1i}(t) = b_{1i}t$ and $f_{2i}(t) = b_{2i}t$ for $i = 2, 3$.
- Axis 1 linear; Axes 2 and 3 quadratic: $f_{11}(t) = b_{11}t$ and $f_{21}(t) = b_{21}t$; and $f_{1i}(t) = b_{1i}t$ and $f_{2i}(t) = b_{2i}t + b_{3i}t^2$ for $i = 2, 3$.
- All axes quadratic: $f_{1i}(t) = b_{1i}t$ and $f_{2i}(t) = b_{2i}t + b_{3i}t^2$ for all i.

The pseudo-loglikelihood (ℓ) and AICc for each model are shown in Table 3.16.

Model	ℓ	m	N	AIC	AICc
All axes linear	-825.751	6	1500	1663.50	1663.54
One quadratic, 2 linear	-824.815	7	1500	1663.63	1663.69
Two quadratic, 1 linear	-823.942	8	1500	1663.88	1663.96
All axes quadratic	-822.989	9	1500	1663.98	1664.08

Table 3.16: Results from model selection

This shows that the first option, all axes linear, is the best fitting (which is not surprising, as this was the generating form for the data). The inclusion of the quadratic terms has resulted in over-fitting, something which the AIC (and AICc) penalise (Sakamoto et al. (1989)).

The model selection is now demonstrated on some real data - the deodorant example from Section 3.4. The model previously fitted used a quadratic axis to represent each attribute. The effect of the quadratic term on the representation of the attributes 1 and 9 on products A and B was highlighted.

Therefore, as an example, the configuration is refitted with i) both these axes linear, ii) Attribute 9 linear, and iii) Attribute 1 linear, to see which configuration is best.

Model	ℓ	m	N	AIC	AICc
All quadratic	-771.117	27	1350	1596.234	1597.296
i) Attributes 1 and 9 linear	-775.654	25	1350	1601.308	1602.214
ii) Attribute 9 linear	-774.119	26	1350	1600.238	1601.221
iii) Attribute 1 linear	773.243	26	1350	1598.486	1599.469

Table 3.17: Results from model selection on Deodorant data

Table 3.17 shows the various AICc results, with the all quadratic option being the best fitting. This process can be repeated across the attributes. Thus, by trying the various models for the axes, it can be seen how the optimal configuration can be found.

Chapter 4

Dynamic Multidimensional Scaling

4.1 Introduction

Multidimensional Scaling, as dealt with so far in this thesis, has produced a map showing the relationships between objects at a set, static, time point. However, often measures on objects are recorded over time, and it is of interest to see how the relationships change. For example, consumers might be asked for their opinions on products over several weeks. One possible solution is to produce a separate map for each time point, and attempt to follow each object as it is depicted in each map. However, a more elegant solution would be to produce one map which shows, as trajectories, how the objects move relative to each other. This overcomes one of the main problems with the separate maps - their not being fixed in scale, location or orientation. Even if generalized Procrustes analysis was applied, there will still be deviations between each map that arise through the mapping process.

This Chapter develops a methodology for such a process, called Dynamic Multidimensional Scaling (DMDS). Figure 4.1 shows a schematic of three MDS maps representing objects measured at three time points, $(t = 0, 1, 2)$. It is difficult to understand how each object is changing. Figure 4.2 shows a schematic of a DMDS map on the same data. Now it is relatively easy to see what is happening - for example, the objects in the lower left hand quadrant are becoming more similar over time, as their trajectories are moving closer together. The length of each trajectory can also be thought of as an indication of how much variability over time there is for an object.

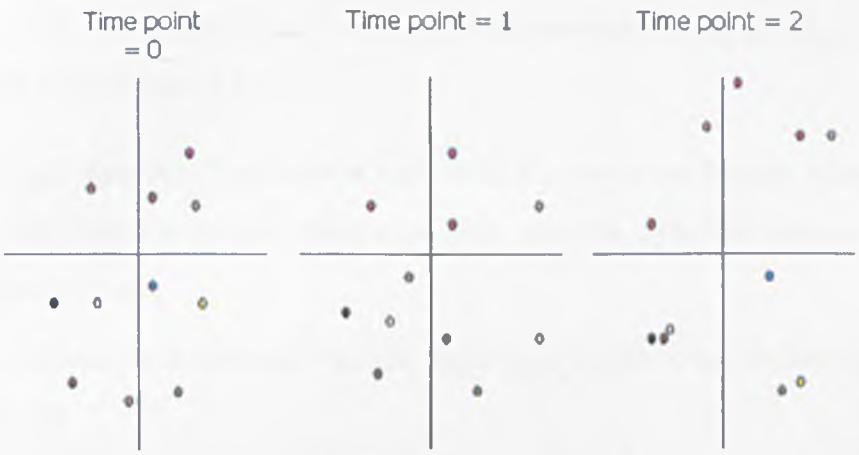

Figure 4.1: Schematic of three MDS maps showing objects at three different time points

4.2 Previous work

Let there be n objects measured at T successive time points, where δ_{ij}^t is the dissimilarity between object i and j at time point t, $i, j = 1, \ldots, n, t = 1, \ldots, T$. The aim is to produce a configuration of nT points in a space, where each object is represented T times, once for each of the time points.

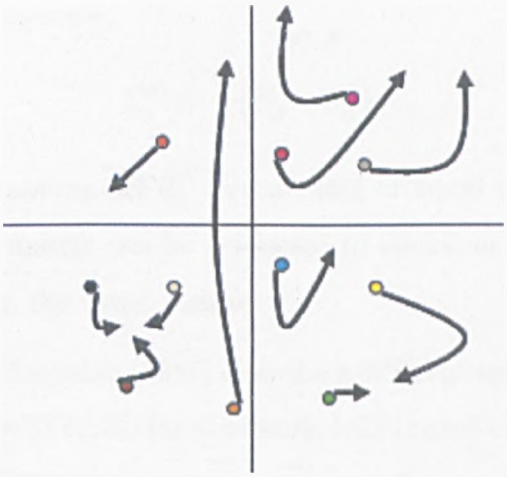

Figure 4.2: Schematic of a Dynamic Multidimensional Scaling map, based on data from Figure 4.1

It is hoped that the T points for each object are not too distant from each other, and that by plotting their path over time the dynamic nature of the data can be seen.

One approach is to place all the dissimilarities into a super-dissimilarity matrix, D,

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}_{11} & \mathbf{D}_{12} & \cdots & \mathbf{D}_{1T} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{D}_{T1} & \mathbf{D}_{T2} & \cdots & \mathbf{D}_{TT} \end{bmatrix},$$

where $\mathbf{D}_{tt} = \left(\delta_{ij}^{t}\right)$, the dissimilarity matrix formed from the dissimilarities collected at the tth time period.

In addition the matrix $\mathbf{D}_{tt'} = \left(\delta_{ij}^{t,t'}\right)$ has to be specified, where $\delta_{ij}^{t,t'}$ is the dissimilarity of object i at time point t with object j at time point t', $t \neq t'$. Some information may be available from which these cross time period dissimilarities can be found. Alternatively they could be constructed

from (δ_{ij}^t), by, for example,

$$\delta_{ij}^{t,t'} = \frac{1}{2}\left(\delta_{ij}^t + \delta_{ij}^{t'}\right).$$

Another way is to assume all $\delta_{ij}^{t,t'}$ are missing or equal to zero. However it is constructed, the matrix can be subjected to metric or nonmetric multidimensional scaling in the usual manner.

Ambrosi and Hansohm (1987) describe a different approach for analysing such data. They use STRESS for nonmetric MDS based on the dissimilatities for time period t defined by

$$STRESS^t = \frac{\sum_{i<j}\left(\delta_{ij}^t - \hat{d}_{ij}^t\right)^2}{\sum_{i<j}\left(\hat{d}_{ij}^t - \bar{d}^t\right)^2},$$

where

$$\bar{d}^t = \frac{2}{n(n-1)}\sum_{i<j}\hat{d}_{ij}^t$$

and $\hat{d}_{ij}^t$ is the estimated distance between the points representing objects i and j at time point t. The combined STRESS for the T time points can be chosen as either

$$S = \frac{\sum_{t=1}^{T}\sum_{i<j}\left(\delta_{ij}^t - \hat{d}_{ij}^t\right)^2}{\sum_{t=1}^{T}\sum_{i<j}\left(\hat{d}_{ij}^t - \bar{d}^t\right)^2}$$

or

$$S = \sum_{t=1}^{T} STRESS^t.$$

This overall STRESS is minimised with the penalty that, in the resulting configuration, the T points that represent each object tend to be close to each other. This is achieved by using a penalty function such as

$$U = \sum_{t=1}^{T-1}\sum_{i=1}^{n}\sum_{m=1}^{p}\left(x_{im}^{t+1} - x_{im}^t\right)^2,$$

where $\mathbf{x}_i^t = (x_{i1}^t, \ldots, x_{ip}^t)$ are the coordinates representing object i at time period t. A configuration is then chosen that minimises

$$STRESS_\epsilon = STRESS + \epsilon U, \quad \epsilon > 0,$$

where ϵ is a chosen constant $<< 1$. Minimising the STRESS and the penalty function U is then a compromise which depends on the value of ϵ, which in turn will depend on the importance placed on the requirement that the T points representing an object are near to each other.

One problem with this method is that there is still the possibility that the solution will not give an understandable representation of how the objects are moving over time. The methodology described in the next section shows how the points representing each object can be restricted to lie on a curved trajectory. This is the process shown in Figure 4.2.

4.3 Dynamic Multidimensional Scaling

Let $\mathbf{X}$ represent the original data set, and use a distance measure (for example Euclidean distance) to calculate the proximity (δ_{ij}^t) between each and every pair of objects i and j, at each time point t, for all $i, j = 1, \ldots, N$ and $t = 1, \ldots, T$.

Now the trajectories for each object $i = 1, \ldots, N$ are defined as

$$x_{1i}^t = x_{1i}^1 + f_{1i}(t), \quad \text{and} \quad x_{2i}^t = x_{2i}^1 + f_{2i}(t), \quad t = 2, \ldots, T \qquad (4.1)$$

where (x_{1i}^t, x_{2i}^t) are the coordinates of object i at time point t, with (x_{1i}^1, x_{2i}^1) being the initial starting coordinates for object i at time point $t = 1$, and $f_{1i}(t)$ and $f_{2i}(t)$ are the defining functions for the trajectory of object i.

A STRESS value is defined as

$$\text{STRESS} = \frac{\sum_{t=1}^{T}\sum_{t'=1}^{T}\sum_{i<j}^{N}\left(\delta_{ij}^{t} - \hat{d}_{ij}^{t'}\right)^2}{T\sum_{t=1}^{T}\sum_{i<j}^{N}\left(\delta_{ij}^{t}\right)^2} \tag{4.2}$$

where $\hat{d}_{ij}^{t'} = \left(\left(x_{1i}^{t'} - x_{1j}^{t'}\right)^2 - \left(x_{2i}^{t'} - x_{2j}^{t'}\right)^2\right)^{\frac{1}{2}}$, i.e. the Euclidean distance between the points at time t', which have been forced onto the trajectory defined in equation 4.1.

The summation is over t and t' in order to take into account the inter-time point distances, as well as the intra-time point distances. The inter-time point distances are included so as to make the configuration more robust. If they are not included, the process is such that the distances between time points for each object can be arbitrarily large. The inclusion of the cross-time points prevents this from happening and produces a 'tighter' configuration. In addition, a weighting could be included, such as

$$\omega^{tt'}\left(\delta_{ij}^{t} - \hat{d}_{ij}^{t'}\right)^2.$$

This would allow different time points to have different impacts. For instance if $\omega^{tt'} = 1$ for $|t - t'| = 1$ and zero otherwise, then STRESS is not influenced by $\hat{d}_{ij}$'s far apart timewise. When $t = t'$ then the weighting should always be one.

The STRESS is minimised with respect to x_{1i}^{1}, x_{2i}^{1} and the parameters in $f_{1i}(t)$ and $f_{2i}(t)$ for all $i = 1, \ldots, N$ iteratively by the Nelder-Mead simplex method found in PROC IML in the SAS software package (Nelder and Mead (1965) and Powell (1992)).

4.4 Examples and Applications

In this section, the methodology is demonstrated on simulated and real data sets.

4.4.1 Simulated data

This section details the application of Dynamic MDS to simulated data. As the data has been user-generated, then it is known what the results should be, which can give a measure of how well the method can represent the data.

An initial 2-dimensional starting configuration of ten objects was generated, with coordinates (x_{1i}^1, x_{2i}^1) in both dimensions randomly selected from a uniform distribution between -5 and +5. Next, the trajectories were simulated as $(x_{1i}^t = x_{1i}^1 + \alpha_{1i}t + \alpha_{2i}t^2, x_{1i}^t = x_{2i}^1 + \beta_{1i}t + \beta_{2i}t^2)$, where the parameters were selected from a uniform distribution between -0.5 and +0.5. The simulated data can be seen in Figure 4.3, and the starting configuration and parameters in Table 4.1. The letters identifying the objects are at the start of the trajectories.

As can be seen in Figure 4.3, the products start-off quite similar, and as time progresses they tend to move further apart.

At time points $t = 1, \ldots, 6$, the locations of the trajectories were calculated, and the Euclidean distance between each location within a time point calculated. This led to 6 individual distance matrices, which form the input data for the method.

Carrying out the method on the data results in the configuration shown in Figure 4.4. This solution has a STRESS of 0.0166.

Object (i)	x^1_{1i}	x^1_{2i}	α_{1i}	α_{2i}	β_{1i}	β_{2i}
A	0.241	-2.016	-0.303	-0.019	-0.476	-0.199
B	3.620	0.248	-0.179	-0.166	-0.484	0.000
C	-2.329	3.740	0.153	0.075	0.184	0.069
D	-0.760	0.221	-0.453	0.091	0.134	-0.156
E	0.483	1.055	-0.265	0.045	-0.467	0.080
F	2.381	-0.037	0.482	-0.151	-0.177	-0.162
G	-4.376	-0.178	0.027	0.162	-0.314	-0.119
H	0.634	3.964	-0.101	0.135	-0.117	-0.090
I	-0.619	-2.660	0.122	-0.087	-0.286	-0.044
J	-2.775	3.943	0.102	0.018	0.079	-0.094

Table 4.1: Starting configuration and parameters for simulated data

The reason why the STRESS is not zero is that the Dynamic MDS algorithm involves a cross-time comparison (the t' in Equation 4.2) which impacts on the fit. The cross-time term was not factored into the data simulation process.

Despite this, the configuration shown in Figure 4.4 matches that of Figure 4.3 relatively well. However, the configuration has been rotated 90° anti-clockwise. In other words, as with usual MDS, Dynamic MDS does not result in a unique configuration.

4.4.2 Investigating hair styles over time

Twelve panellists were involved in a study to investigate how their hair style changed over time. Each panellist had their hair professionally styled, and was then asked to complete a daily questionnaire over a week.

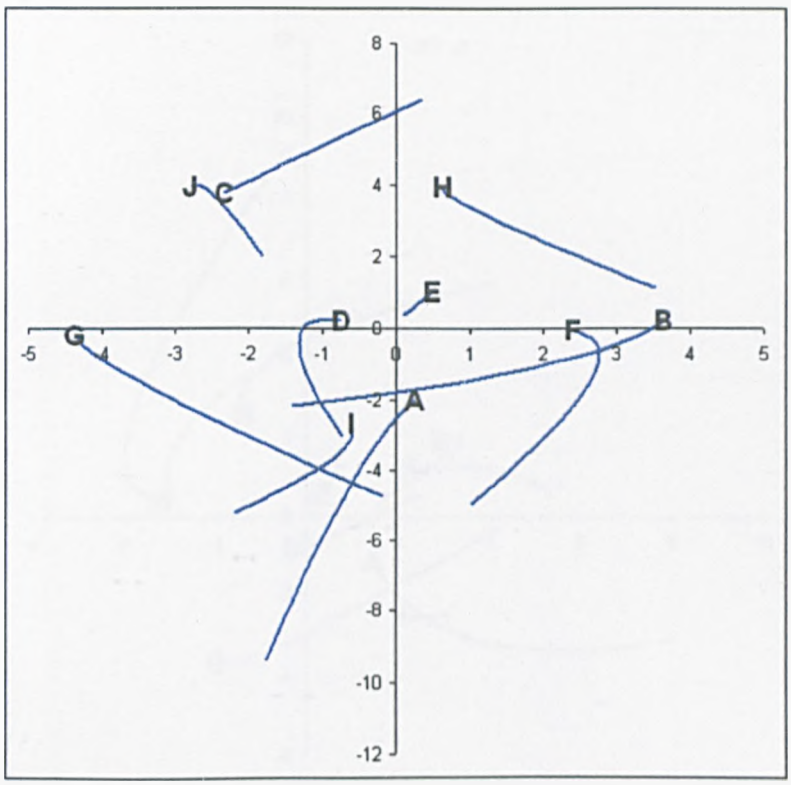

Figure 4.3: Simulated data to be recreated by Dynamic MDS. Trajectories start at the letter.

Figure 4.5 shows the Dynamic MDS plot produced from the data. The STRESS value of 0.475 indicates that this is a good representation of the data. Looking at the plot shows that the majority of panellists are moving in the same direction towards the bottom of the graph - i.e. their opinions of their hair are changing in the same way over time. Investigation of the raw data reveals that these panellists are perceiving a worsening of their hair style over time. Some panellists, however, are moving in the opposite direction, and these are found to be happy with their style, and also score highly on style retention.

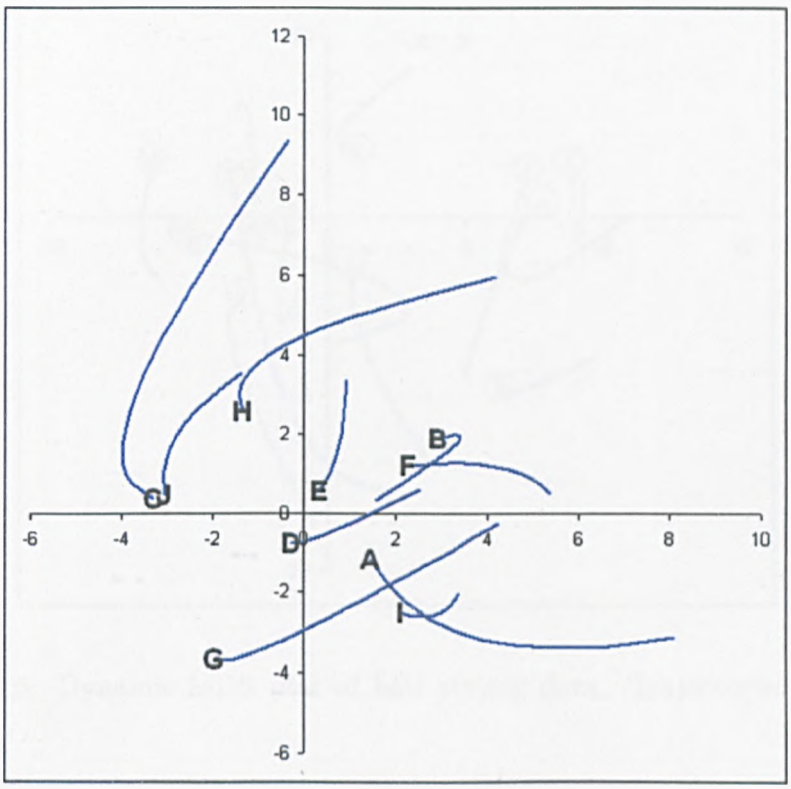

Figure 4.4: Results of Dynamic MDS on simulated data. Trajectories start at the letter.

4.4.3 Spray characteristics of spray cans with different fill weights

The defining form of the trajectories is not limited to time - any relevant axis can be used. In this example, Dynamic MDS is used to investigate how the spray characteristics of nine deodorants change over varying levels of how full the can is. Nine deodorants (A,B,C,D,E,F,G,H and I) had several spray characteristics measured at each of five levels of can fill (20%, 40%, 60%, 80% and 100%). The configuration generated by Dynamic MDS, which had a STESS of 0.0538, is shown in Figure 4.6.

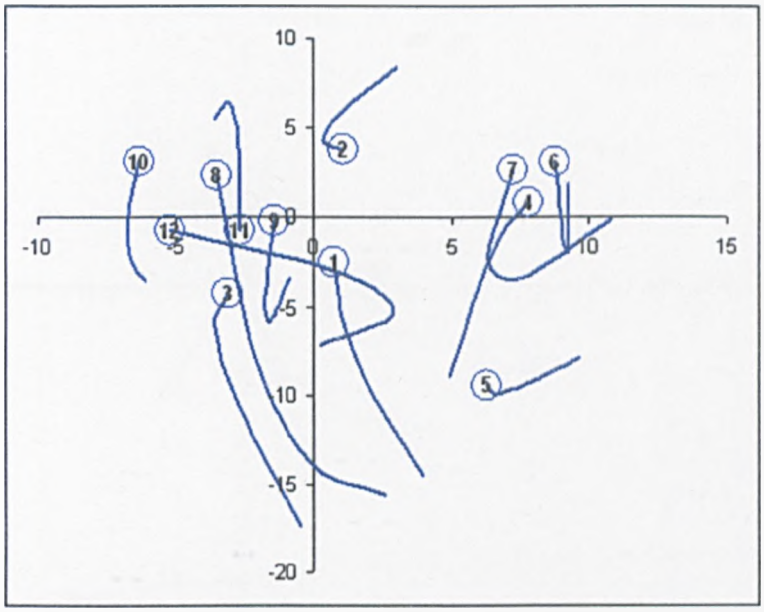

Figure 4.5: Dynamic MDS plot of hair styling data. Trajectories start at label

When they are nearly empty, many of the deodorants have similar characteristics - for example, products F, G and H all start their trajectories in the centre of the map. With increasing fill, the characteristics start to differ, as can be seen with the trajectories radiating outwards. Product C has a unique set of characteristics, possibly due to its unique formulation.

4.5 Goodness of Fit

Section 4.4 showed that the methodology can recreate simulated data which contains no noise. It has also been seen that the STRESS value is a measure of how well the configuration represents the data. This section deals with investigating what happens when noise is added to the simulated data and how this impacts on the STRESS.

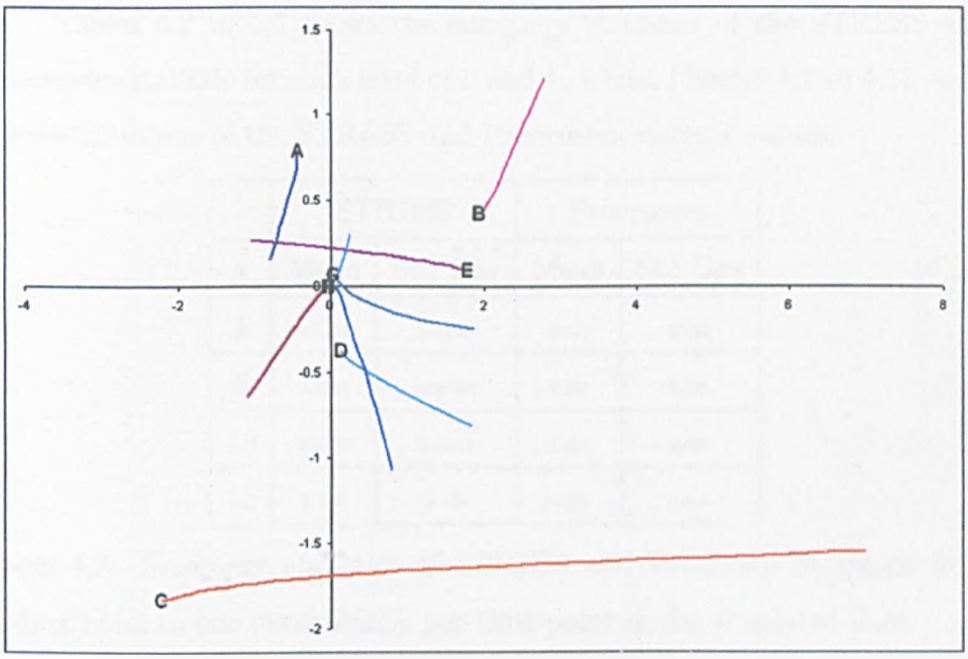

Figure 4.6: Dynamic MDS plot of spray characteristics data. Trajectories start at label

Using the simulated data, noise was added to the dissimilarities. Initially, one δ_{ij} was chosen at random for each time point, and random error added to it. The value of the error was selected from a uniform distribution between $-k$ and $+k$. The value of k was altered to investigate the effect of increasing levels of noise, with k=1,5,10 and 15 being the selected values. The maximum level of k equates almost to the maximum δ_{ij} of 16.8. In addition, the number of altered dissimilarities per time point was varied, with r=1,5,10,20,25 being the selected values. If any altered dissimilarity became less than zero, it was set to be zero. For each combination of r and k, the method was run 1000 times, and the STRESS recorded. In addition, a Procrustes statistic was calculated to see how closely the method fitted the altered data to the original data.

Tables 4.2 to 4.6 shows the summary statistics of the STRESS and Procrustes statistic for each level of r and k, whilst Figures 4.7 to 4.11 show the distributions of the STRESS and Procrustes statistic values.

	STRESS		Procrustes	
k	Mean	Std Dev	Mean	Std Dev
1	0.0196	0.00335	0.151	0.012
5	0.0351	0.00765	0.151	0.012
10	0.0702	0.0217	0.152	0.016
15	0.109	0.404	0.152	0.016

Table 4.2: Summary statistics of STRESS and Procruste Statistics from adding noise to one dissimilarity per time point in the simulated data

Figure 4.7: Distribution of STRESS values and Procrustes Statistics from adding noise to one dissimilarity per time point in the simulated data

	STRESS		Procrustes	
k	Mean	Std Dev	Mean	Std Dev
1	0.0234	0.00367	0.152	0.012
5	0.0954	0.0123	0.157	0.021
10	0.288	0.434	0.168	0.027
15	0.429	0.630	0.166	0.025

Table 4.3: Summary statistics of STRESS and Procruste Statistics from adding noise to five dissimilarities per time point in the simulated data

Figure 4.8: Distribution of STRESS values and Procrustes Statistics from adding noise to five dissimilarities per time point in the simulated data

	STRESS		Procrustes	
k	Mean	Std Dev	Mean	Std Dev
1	0.0269	0.00389	0.153	0.011
5	0.168	0.0189	0.159	0.020
10	0.423	0.0477	0.166	0.027
15	0.684	0.0693	0.172	0.028

Table 4.4: Summary statistics of STRESS and Procruste Statistics from adding noise to ten dissimilarities per time point in the simulated data

Figure 4.9: Distribution of STRESS values and Procrustes Statistics from adding noise to ten dissimilarities per time point in the simulated data

	STRESS		Procrustes	
k	Mean	Std Dev	Mean	Std Dev
1	0.0330	0.00464	0.153	0.014
5	0.303	0.0269	0.166	0.025
10	0.707	0.0558	0.177	0.028
15	0.995	0.0693	0.184	0.027

Table 4.5: Summary statistics of STRESS and Procruste Statistics from adding noise to fifteen dissimilarities per time point in the simulated data

Figure 4.10: Distribution of STRESS values and Procrustes Statistics from adding noise to fifteen dissimilarities per time point in the simulated data

	STRESS		Procrustes	
k	Mean	Std Dev	Mean	Std Dev
1	0.0365	0.00416	0.150	0.013
5	0.364	0.0308	0.167	0.014
10	1.274	.0625	0.177	0.027
15	10.98	0.778	0.192	0.025

Table 4.6: Summary statistics of STRESS and Procruste Statistics from adding noise to twenty dissimilarities per time point in the simulated data

Figure 4.11: Distribution of STRESS values and Procrustes Statistics from adding noise to twenty dissimilarities per time point in the simulated data

With increasing amounts of noise, there is, as expected, an increase in the STRESS and Procrustes statistic. However, it seems that the methodology is fairly robust, and even altering about half the dissimilarities by up to five units has very little impact on the fit statistics.

4.6 Model Selection

As the STRESS value is a measure of how well the configuration represents the data, then it seems obvious to use it in selecting the 'best' configuration. Different configurations, using different functions to describe the trajectories, can be fitted, and the one with the lowest STRESS selected. However, care must be taken not to overfit the data - for example a quadratic term in the function would be expected to represent the data better than just the linear term. Thus, some penalty term should be included in the diagnostic to penalise over-fitting - in a similar way to the AIC of Section 3.6.

However, as can be seen in Figure 4.6, even when quadratic functions are used, the parameter for the quadratic term is often very close to zero when it is not needed to increase the fit. Whether this is true in all cases remains to be seen. Compare though the trajectories shown in Figures 4.5 and 4.6.

Chapter 5

Summary, Conclusions and Future Work

Multidimensional Scaling has been demonstrated to have a wide variety of applications, where multivariate data can be visualised in terms of proximities between objects. Within the fields of sensory and consumer science, MDS is recognised as an important method of data visualisation and analysis. Because of its importance, a decision was made to develop adaptations to the methodology within this thesis.

The addition of curved axes to MDS maps was an attempt to aid interpretation of the maps. By overlaying an axis that represents an attribute, it is possible to see how the objects score on that attribute. Thus, the axis guides the researcher by highlighting on the map how different attributes are changing. It is also possible when several axes are overlaid to determine which attributes are related. In a way, this resembles the biplot - attributes are represented by axes (or vectors), and objects by points, and it is possible to determine the relationships between attributes and objects.

The methodology developed within this thesis allows for axes to be fitted to MDS plots of any dimensionality. The axes can be described by a set of differentiable functions. However, this leads to a problem - that of which is the best function to describe an axis. Simple linear terms would result in a straight line, whilst the addition of extra terms such as quadratics intuitively lead to 'better' fitting axes. However, one area of concern is how to define 'better', whilst there is also the problem of overfitting. As with linear regression, the more terms that are added to an equation, the more the model is said to be overfitted, with the potential of just modelling random flucuations in the data. This is also a problem with the methodology. Whilst an attempt was made in the thesis to look at measuring the goodness-of-fit, using cross-validation methods, there remains more work that can be done in this area. For example, different metrics could be developed to determine how well a function represents the attribute. Or other methods could be used for comparing the axes, and ultimately solving the model selection problem.

Returning to the dimensionality of the solution, this can also cause a problem. Using the map example of Section 1.2 as an example - here the map of England is easily shown in two dimensions. However, suppose the input data was flight times between major cities of the world, with examples from all the continents. Now a two-dimensional map would not be satisfactory, as the points representing the cities would be expected to lie on a sphere, representing the shape of the globe. The axes representing longitude and latitude would now be expected to be circles shown in their relevant plane. However, would this three-dimensional representation be the 'best' representation?

Finally, work is needed to look at the loss function - what is an acceptable value?

In Chapter 3, the methodology for displaying multivariate paired com-

parison data was developed. Again, as with the fitting of curved axes, one area where more research is needed is model selection - in other words, which functions should be used to define the axes. An Akaike's Information Criterion approach was suggested, but more work is needed to determine if this is the best approach. Additionally, dimensionality of the solution needs to be factored into the goodness of fit measures, and further investigation into the behaviour of the pseudo-likelihood is needed.

When two MDS-like configurations have been produced from multivariate paired-comparison data, a method is needed for comparing them. It could be possible to compare just the configuration of the points using Generalised Procrustes Analysis. However, this does not take into account the location of the axes. Certain points could be calculated along each axis (for example, when $t = -2, -1, 0, +1, +2$, and these coordinates used within the GPA. However, whether this would be a suitable method remains to be seen.

Concerning the Dynamic Multidimensional Scaling procedure described in Chapter 4, one major adaptation that could be made is the use of time series models for describing the axes instead of the functions already described. A simple example would be

$$x_{1i}^{t} = x_{1i}^{0} + \alpha x_{1i}^{t-1}$$

and

$$x_{2i}^{t} = x_{2i}^{0} + \beta x_{2i}^{t-1}.$$

The advantage of this would be that the time series axes could be used to incorporating things like seasonality, assuming enough time points had been calculated. However, the major difficulty with such an approach is the conversion of the time series models into Cartesian coordinates for plotting.

Another problem that could arise with the Dynamic Multidimensional Scaling is missing data points. If for example, objects are censored (for example, a panellist did not complete all the questionnaires, but instead dropped out half way through the study), or simply just one or two time points are missing for an object, then this will have an impact on the STRESS calculations. There is no way of incorporating the missing data into the inter-time point calculations. Also, a similar problem is possible if the observations are not made at the same time points. A method needs to be solved for bringing such data into the calculation of the STRESS.

As mentioned in Section 4.6, there is currently no method for model selection. Whilst the STRESS provides a metric for determining how well the configuration is fitting the data, this does not take into account any possible over-fitting from the functions. As mentioned earlier a penalty function needs to be included. Thus a metric would be calculated that would enable comparisons of the different configurations, and thus selection of the 'best'.

Because Dynamic MDS produces a unique set of functions for each object, could these be used for some form of clustering of the objects. For example, objects which have trajectories that move in the same direction could be said to be behaving similarly over time. Or objects with trajectories that converge are becoming more similar over time could be clustered together.

Finally there is the possibility of combining some of the methods described in the thesis. It is possible that once a configuration has been produced either by Dynamic MDS or the MDS approach to multivariate paired comparison, then description axes could be overlaid by the methodology in Chapter 2. This would enable a better understanding of the output. Alternatively could there be some of way of taking multivariate paired comparison

data that has been generated over time, and producing a Dynamic approach to analysing this?

Appendix A

Classical Scaling

MDS was first developed in work by Eckart and Young (1936) and Young and Householder (1938). Torgerson (1952) used this earlier work to develop the MDS method known as *classical scaling.* Classical scaling is motivated by the need mentioned in Section 1.3 - to calculate a configuration $\mathbf{X}$, in p-space which has inter-point distances given in $\mathbf{\Delta}$, where $\mathbf{X} = (x_{ir})$, a configuration matrix of n points in p-space where x_{ir} denotes the r^{th} coordinate of point i and Δ is a matrix of inter-point Euclidean distances. Before explaining classical scaling, we need some definitions.

Definition A.0.1.

A matrix Δ *of dissimilarities* d_{ij} *is* Euclidean *if there exists a dimension* k *and a set of* n *points* $x_1, x_2, \ldots, x_n \in \mathbb{R}^k$ *such that* $d_{ij}^2 = (x_i - x_j)^T (x_i - x_j)$ $(i, j = 1, 2, \ldots, n)$.

Definition A.0.2.

The matrix $\mathbf{A} = (a_{ij})$ *is defined by* $a_{ij} = -\frac{1}{2} d_{ij}^2$. *The matrix* $\mathbf{B}_\Delta = (b_{ij})$ *is defined by* $\mathbf{B}_\Delta = \mathbf{CAC}$ *where* $\mathbf{C} = \mathbf{I_n} - \frac{1}{n} \mathbf{1_n} \mathbf{1_n^T}$.

Note that the above definition implies that $b_{ij} = a_{ij} - \overline{a}_{i.} - \overline{a}_{.j} + \overline{a}_{..}$ where $\overline{a}_{i.} = \frac{1}{n}\sum_{j=1}^{n} a_{ij}$, $\overline{a}_{.j} = \frac{1}{n}\sum_{i=1}^{n} a_{ij}$, $\overline{a}_{..} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}$.

We can now answer the classical scaling question by stating what Jobson (1992) calls the *fundamental theorem of MDS.*

Theorem A.0.1.

A dissimilarity matrix Δ *is Euclidean* $\Longleftrightarrow$ Δ *is such that* $\mathbf{B}_\Delta = \mathbf{CAC}$ *is positive semi-definite.*

Proof. $\Rightarrow$

Let Δ be Euclidean so there exist $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathbb{R}^k$ such that $d_{ij}^2 = (\mathbf{x}_i - \mathbf{x}_j)^T (\mathbf{x}_i - \mathbf{x}_j)$ $(i, j = 1, 2, \ldots, n)$.

Then $a_{ij} = -\frac{1}{2}d_{ij}^2 = -\frac{1}{2}\left(\mathbf{x}_i^T\mathbf{x}_i + \mathbf{x}_j^T\mathbf{x}_j - 2\mathbf{x}_i^T\mathbf{x}_j\right)$.

Let $\overline{\mathbf{x}^T\mathbf{x}} = \frac{1}{n}\sum_{r=1}^{n} \mathbf{x}_r^T\mathbf{x}_r$ and $\overline{\mathbf{x}} = \frac{1}{n}\sum_{r=1}^{n} \mathbf{x}_r$.

Then $\overline{a}_{i.} = -\frac{1}{2}\left(\mathbf{x}_i^T + \overline{\mathbf{x}^T\mathbf{x}} - 2\mathbf{x}_i^T\overline{\mathbf{x}}\right)$, $\overline{a}_{.j} = -\frac{1}{2}\left(\overline{\mathbf{x}^T\mathbf{x}} + \mathbf{x}_j^T\mathbf{x}_j - 2\overline{\mathbf{x}}^T\mathbf{x}_j\right)$,

$\overline{a}_{..} = -\frac{1}{2}\left(2\overline{\mathbf{x}^T\mathbf{x}} - 2\overline{\mathbf{x}}^T\overline{\mathbf{x}}\right)$,

so that $b_{ij} = a_{ij} - \overline{a}_{i.} - \overline{a}_{.i} + \overline{a}_{..} = \mathbf{x}_i^T\mathbf{x}_j - \mathbf{x}_i^T\overline{\mathbf{x}} - \overline{\mathbf{x}}^T\mathbf{x}_j + \overline{\mathbf{x}}^T\overline{\mathbf{x}}$

$= (\mathbf{x}_i - \overline{\mathbf{x}})^T (\mathbf{x}_j - \overline{\mathbf{x}})$.

Now let $\mathbf{X} = (\mathbf{x}_1\, \mathbf{x}_2 \ldots \mathbf{x}_n)^T$ and consider

$\mathbf{X} - \mathbf{1}_n\overline{\mathbf{x}}^T = \left((\mathbf{x}_1 - \overline{\mathbf{x}})(\mathbf{x}_2 - \overline{\mathbf{x}})\ldots(\mathbf{x}_n - \overline{\mathbf{x}})\right)^T$ so that

$\left(\mathbf{X} - \mathbf{1}_n\overline{\mathbf{x}}^T\right)\left(\mathbf{X} - \mathbf{1}_n\overline{\mathbf{x}}^T\right)^T =$

$$\begin{pmatrix} (\mathbf{x}_1 - \overline{\mathbf{x}})^T(\mathbf{x}_1 - \overline{\mathbf{x}}) & \ldots & (\mathbf{x}_1 - \overline{\mathbf{x}})^T(\mathbf{x}_n - \overline{\mathbf{x}}) \\ & \vdots & \\ (\mathbf{x}_n - \overline{\mathbf{x}})^T(\mathbf{x}_1 - \overline{\mathbf{x}}) & \ldots & (\mathbf{x}_n - \overline{\mathbf{x}})^T(\mathbf{x}_n - \overline{\mathbf{x}}) \end{pmatrix}$$

$$= \begin{pmatrix} b_{11} & \dots & b_{1n} \\ \vdots & & \\ b_{n1} & \dots & b_{nn} \end{pmatrix}$$

i.e. $\mathbf{B}_{\Delta} = \left(\mathbf{X} - \mathbf{1}_n\bar{\mathbf{x}}^T\right)\left(\mathbf{X} - \mathbf{1}_n\bar{\mathbf{x}}^T\right)^T$.

Hence where $\mathbf{z} \in \mathbb{R}^n$, $\mathbf{z}^T\mathbf{B}_{\Delta}\mathbf{z} = \mathbf{z}^T\left(\mathbf{X} - \mathbf{1}_n\bar{\mathbf{x}}^T\right)\left(\mathbf{X} - \mathbf{1}_n\bar{\mathbf{x}}^T\right)^T\mathbf{z} = \mathbf{y}^T\mathbf{y} \geq 0$ where $\mathbf{y} = \left(\mathbf{X} - \mathbf{1}_n\bar{\mathbf{x}}^T\right)$ and thus $\Delta$is positive semi-definite. □

Proof. $\Leftarrow$

Let Δ be such that Δ is positive semi-definite. $\mathbf{B}_{\Delta}{}^T = (\mathbf{CAC})^T$

$= \mathbf{C}^T\mathbf{A}\mathbf{C}^T = \mathbf{CAC}$ so Δ is symmetric. Δ has n non-negative eigenvalues since it is positive semi-definite. Suppose k of these are positive and the rest zero, so that the eigenvalues are listed as

$$\lambda_1 \geq \lambda_2 \geq \dots \lambda_k > \lambda_{k+1} = \dots = \lambda_n = 0$$

with corresponding normalised eigenvectors

$$\mathbf{l}_1, \mathbf{l}_2, \dots, \mathbf{l}_k, \mathbf{l}_{k+1}, \dots, \mathbf{l}_n.$$

Define the matrices $\Lambda_1 = \text{diag}\,(\lambda_1, \lambda_2, \dots, \lambda_k)$, $\Lambda_2 = \text{diag}\,(\lambda_{k+1}, \lambda_{k+2}, \dots, \lambda_n) = \text{diag}\,(0, 0, \dots, 0)$, $\mathbf{L}_1 = (\mathbf{l}_1, \mathbf{l}_2, \dots, \mathbf{l}_k)$ and $\mathbf{L}_2 = (\mathbf{l}_{k+1}, \dots, \mathbf{l}_n)$.

The spectral decomposition of Δ is then

$$\mathbf{B}_{\Delta} = \mathbf{L}\Lambda\mathbf{L} = \begin{pmatrix} \mathbf{L}_1 & \mathbf{L}_2 \end{pmatrix}\begin{pmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{pmatrix}\begin{pmatrix} \mathbf{L}_1^T \\ \mathbf{L}_2^T \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{L}_1 & \mathbf{L}_2 \end{pmatrix}\begin{pmatrix} \Lambda_1 & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} \mathbf{L}_1^T \\ \mathbf{L}_2^T \end{pmatrix} = \mathbf{L}_1\Lambda_1\mathbf{L}_1^T$$

$$= \mathbf{L}_1\Lambda_1^{1/2}\Lambda_1^{1/2}\mathbf{L}_1^T = \mathbf{L}_1\Lambda_1^{1/2}\left(\mathbf{L}_1\Lambda_1^{1/2}\right)^T = \mathbf{X}\mathbf{X}^T \text{ (say)}.$$

We have obtained a set of points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathbb{R}^k$ which are contained in the matrix $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)^T$ by having $\mathbf{\Delta}$ be such that $\mathbf{\Delta}$ is positive semi-definite. We will show that the inter-point distances in this configuration match $\mathbf{\Delta}$ so that the result is proved.

Now $\mathbf{\Delta} = \mathbf{X}\mathbf{X}^T$ so that $b_{ij} = \mathbf{x}_i^T \mathbf{x}_j$, and therefore

$$(\mathbf{x}_i - \mathbf{x}_j)^T (\mathbf{x}_i - \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_i - 2\mathbf{x}_i^T \mathbf{x}_j = b_{ii} - 2b_{ij} + b_{jj}$$

$$= (a_{ii} - \bar{a}_{i.} - \bar{a}_{.j} + \bar{a}_{..}) - 2(a_{ij} - \bar{a}_{i.} - \bar{a}_{.j} + \bar{a}_{..}) + (a_{jj} - \bar{a}_{i.} - \bar{a}_{.j} + \bar{a}_{..})$$

$$= a_{ii} - 2a_{ij} + a_{jj} = -2a_{ij} = \delta_{ij}^2 \quad \square$$

The latter proof is the important one, suggesting as it does, an algorithm for producting a classical scaling MDS configuration given $\mathbf{\Delta}$:

1. Construct $\mathbf{A}$ from $\mathbf{\Delta}$.
2. Construct $\mathbf{B}_\Delta = \mathbf{CAC}$.
3. Examine the largest k positive eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k$ of $\mathbf{B}_\Delta$ and choose a dimensionality $p \leq k$ in which to display the MDS configuration.
4. Put the normalised eigenvectors $\mathbf{l}_1, \mathbf{l}_2, \ldots, \mathbf{l}_p$ corresponding to the p largest eigenvectors $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ into the matrix $\mathbf{L}_p = (\mathbf{l}_1, \mathbf{l}_2, \ldots, \mathbf{l}_p)$ and form $\Lambda_p = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$.
5. The MDS configuration matrix $\mathbf{X}$ of order n x p (n points in p-space) is given by $\mathbf{X} = \mathbf{L}_p \Lambda_p^{\frac{1}{2}}$.

Note that producing an MDS solution in $p < k$ dimensions will render the configuration approximate rather than exact as is the case when $p = k$.

Should we have a matrix of dissimilarities $\mathbf{\Delta}$ which produces a matrix $\mathbf{B}$ which is not positive semi-definite (negative eigenvalues $\Rightarrow$ $\mathbf{B}$ not positive semi-definite) then there are two options. The first is to discard the negative eigenvalues and execute the algorithm as usual, which gives an approximate solution using the remaining positive eigenvalues. Alternatively the solution to the *additive constant problem* can be employed. Cailliez (1983) produced value c^*, such that when $c > c^*$ is added to all off-diagonal dissimilarities δ_{rs} in $\mathbf{\Delta}$, the matrix $\mathbf{B}_{\mathbf{\Delta}}$ becomes positive semi-definite. The value c^* is found to be the largest eigenvalue of the matrix

$$\begin{pmatrix} 0 & \mathbf{B}_{\mathbf{\Delta}} \\ -\mathbf{I}_n & -4\mathbf{B}_{\mathbf{\Delta}_c} \end{pmatrix}$$

where $\mathbf{B}_{\mathbf{\Delta}_c}$ is formed from square roots of the original dissimilarities.

The final important consideration in classical scaling is that of choosing the appropriate number of dimensions in which to produce the MDS configuration. If the matrix $\mathbf{B}$ is positive semi-definite we can do this by plotting $\sum_{i=1}^{p} \lambda_i / \sum_{i=1}^{n} \lambda_i$ against $p = 1, 2, \ldots, n$, whereas if $\mathbf{B}$ is not positive semi-definite we plot $\sum_{i=1}^{p} \lambda_i / \sum_{i=1}^{n} |\lambda_i|$ or $\sum_{i=1}^{p} \lambda_i / \sum_{i=1}^{n} \lambda_i^2$ or $\sum_{i=1}^{p} \lambda_i / \sum_{i=1}^{n}$(positive eigenvalues) instead.

Appendix B

Non-metric Scaling

The need to provide an explicit function f mapping dissimilarities into Euclidean distances was first relaxed in work by Shepard (1962a). Instead, the only requirement was that f be some unspecified monotonic function, which enabled the finding of (i) the smallest dimensionality for the Euclidean space used to display the MDS configuration with iner-point distances monotonically related to the initial dissimilarities, (ii) the coordinates of the points in the solution configuration, and (iii) a plot showing the shape of the (initially unspecified) function f.

Tests of the algorithm developed in Shepard (1962a) using both simulated and empirical data can be found in Shepard (1962b). In the discussion it is noted that

> "The tests ... have supported the claim that when the proximity measures are monotonically related to distances in an underlying Euclidean configuration, this configuration can be metrically recovered by an analysis based essentially upon the rank order of the proximity measures alone"

Shepard stresses the advantage of his method over traditional MDS techniques - in particular its applicability to the sorts of data found in the psychological literature, such as matrices of *confusion frequencies*, i.e. matrices of the number of times that stimuli in a set are confused with each other. He also points out that his proposed method is more amenable than earlier methods to coping with generalisations such as missing data.

The foundations of Shepard's method were built on by Kruskal (1964a). Kruskal describes *"a technique for multidimensional scaling, similar to Shepard's, which arose from attempts to improve and perfect his ideas"*. His main extension is the introduction of a measure quantifying the quality of the monotonic relationship between a set of dissimilarities and the inter-point distances in a solution configuration in p-space obtained for such a set.

In order to assess the quality of the monotone relationship between a lower triange of dissimilarities δ_{ij} and a set of distances d_{ij}, Kruskal performs a monotone least squares regression of the distances upon the dissimilarities. First, the dissimilarities are arranged in ascending order

$$\delta_{i_1 j_1} < \delta_{i_2 j_2} < \ldots < \delta_{i_M j_M}$$

(where $M = n(n-1)/2$), and then the distances are arranged in the sequence

$$d_{i_1 j_1}, d_{i_2 j_2}, \ldots, d_{i_M j_M}.$$

Next a set of disparities $\hat{d}_{ij}$ are produced according to the following algorithm

1. Let $k = 1$ and set $\hat{d}_{i_1 j_1} = d_{i_1 j_1}$.

2. Increase k by 1.

3. If $d_{i_k j_k} > d_{i_{k-1} j_{k-1}}$ then set $\hat{d}_{i_k j_k} = d_{i_k j_k}$, otherwise let t be the number of distances after $d_{i_{k-1} j_{k-1}}$, which are in the decreasing sequence whose

first term is $d_{i_{k-1}j_{k-1}}$, and set $\hat{d}_{i_{k-1}j_{k-1}}, \ldots, \hat{d}_{i_{k-1+t}j_{k-1+t}} = \frac{1}{t+1}\left(d_{i_{k-1}j_{k-1}} + \ldots + d_{i_{k-1+t}j_{k-1+t}}\right)$ and increase k by $t-1$.

4. If $k \neq M$ then go to step 2.

5. If the disparities are perfectly weakly monotonic then stop, otherwise go to step 1 using the disparities in place of the original distances in order to calculate a new set of disparities.

The way in which disparities are constructed ensures that they are monotonically related to the ordered dissimilarities. A measure of how well the inter-point distances match monotonically the ordered dissimilarities is then given by the *STRESS* $S(\mathbf{X})$ of the p-space configuration $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)^T$. This is defined as

$$S(\mathbf{X}) = \sqrt{\frac{\sum_{i>j}\left(d_{ij} - \hat{d}_{ij}\right)^2}{\sum_{i>j} d_{ij}^2}}$$

The lower the STRESS of a configuration, the better the monotonic relationship, with a STRESS of zero implying a perfect match. Thus the nonmetric technique can be simply stated as the search for a configuration whose STRESS is minimal.

The STRESS can be regarded as simply a function of the np coordinate variables $x_{11}, x_{12}, \ldots, x_{1p}, x_{21}, \ldots, x_{np}$ (via the distances d_{ij}). It is the values of these variables minimising the STRESS which are taken as being the coordinates of the p-space nonmetric MDS configuration. The companion paper Kruskal (1964b) details the algorithm required to find the MDS configuration given a set of dissimilarities δ_{ij}. The algorithm is based on the *method of steepest descent*:

1. Choose an intial configuration of n points in p-space. (Use a good configuration for starting if possible, e.g. the configuration from classical scaling; otherwise an arbitrary starting configuration will suffice).

2. Translate this configuration to have its centroid at the origin.

3. Dilate/contract the configuration so that the root mean square distance of the points from the origin is equal to unity.

4. Calculate the inter-point distances d_{ij} in this standardised configuration.

5. Construct the disparities $\hat{d}_{ij}$.

6. Put $\mathbf{x} = (x_{11}, x_{12}, \ldots, x_{1p}, x_{21}, \ldots, x_{np})^T$ and calculate $\frac{\partial S(\mathbf{X})}{\partial \mathbf{x}}$. If $|\frac{\partial S(\mathbf{X})}{\partial \mathbf{x}}| < \epsilon$, where $\epsilon > 0$ is a small tolerance level, then take $\mathbf{x}$ as the solution configuration and stop. Otherwise continue.

7. Calculate $\alpha_{pres} = 4^{\cos^3\theta} \frac{13g}{10\left(1+s_{five}^5\right)} \alpha_{prev}$ where θ is the angle between gradients on present and previous iterations,
$s_{five} = \min\left(1, \frac{\text{present STRESS}}{\text{STRESS 5 iterations ago}}\right)$, and $g = \min\left(1, \frac{\text{present STRESS}}{\text{previous STRESS}}\right)$.
Note: the initial value of α_{prev} is taken as about $\frac{1}{5}$ when using an arbitrary starting configuration, and should be smaller for a low-STRESS starting configuration. If five iterations have not yet occurred, the quantity 'STRESS five iterations ago' is taken to be the initial 'STRESS' calculated. Similarly we can fix initial values of θ and the quantity 'previous STRESS' on the first iteration.

8. The new configuration vector is given by

$$\mathbf{x}_{new} = \mathbf{x}_{prev} - \alpha_{pres} \frac{\frac{\partial S(\mathbf{X})}{\partial \mathbf{X}}}{\left|\frac{\partial S(\mathbf{X})}{\partial \mathbf{X}}\right|}$$

9. Go to step 2.

There are more recent algorithms for minimising STRESS, such as that of Grönen and Heiser (1996), but the above is the one used to perform non-metric MDS in this thesis. STRESS has been the topic of considerable investigation, and a summary of the work done on the topic can be found in Cox and Cox (2000).

In order to avoid the ubiquitous pitfalls of local minima when minimising a function of many variables, Kruskal recommends starting the steepest descent algorithm from various different starting configurations and bearing in mind that local minima configurations encountered should be rejected anyway unless they have suitably low STRESS.

As in the case of classical scaling, thought needs to be given to how many dimensions are needed for the MDS configuration. The way to determine the appropriate dimensionality is to find configurations in $p = 2, 3, 4, \ldots$ dimensions and then plot a graph of STRESS versus dimension. As p increases, STRESS decreases and the best policy is to choose the smallest p such that the STRESS decrease in moving from p to $p + 1$ dimensions is negligible.

After an MDS configuration has been produced a *Shepard plot* can be constructed by plotting the dissimilarities δ_{ij} and the disparities $\hat{d}_{ij}$ on the vertical axis against the configuration distances on the horizontal axis. Hence, the form of the initially unknown function f relating the dissimilarities and distances can be visualised.

Kruskal (1964a) mentions some important generalisations which could easily be incorporated into his method. The first is when some of the dissimilarities are missing, either deliberately due to constraints on data col-

lection, or accidently. The recommendation here was to simply omit from the STRESS function any terms corresponding to a missing dissimilarity δ_{ij}. The strict requirement that dissimilarities must be symmetric was also relaxed, and the suggestion was to either replace the unequal δ_{ij} and δ_{ji} by $(\delta_{ij} + \delta_{ji})/2$ to achieve symmetry, or generalise STRESS to sum over all dissimilarities rather than just over those δ_{ij} with $i > j$. Another assumption is that there are no *ties* in the data, i.e. dissimilarities are now permitted to be equal. In the presence of such ties, Kruskal provides the *primary approach* whereby the only constraint on the dissimilarities is that

$$\delta_{ij} < \delta_{i'j'} \Rightarrow \hat{d}_{ij} \leq \hat{d}_{i'j'},$$

meaning that $\hat{d}_{ij}$ is not required to be equal to $\hat{d}_{r's'}$, and the *secondary approach* where both

$$\delta_{ij} < \delta_{i'j'} \Rightarrow \hat{d}_{ij} \leq \hat{d}_{i'j'} \quad \text{and} \quad \delta ij = \delta_{i'j'} \Rightarrow \hat{d}_{ij} = \hat{d}_{i'j'}$$

hold.

Another generalisation described is where the distance function d_{ij} used in the STRESS function is not the Euclidean distance, but instead one of the more general *Minkowski c-metrics* d_c whereby for two points $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$ and $\mathbf{x}_j = (x_{j1}, x_{j2}, \ldots, x_{jp})^T$ the inter-point distance is given by

$$d_c(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{r=1}^{p} |x_{ir} - x_{jr}|^c \right)^{\frac{1}{c}}$$

with the Euclidean distance occuring when $c = 2$.

An important distinction between the Euclidean distance and other more general distances is that whilst configurations obtained using Euclidean distance can be rotated, configurations using other distances cannot. Everitt

and Dunn (2001) plot STRESS versus k for MDS solutions in 2-space in order to determine the best Minkowski c-metric to choose.

Finally it should be noted that although nonmetric MDS is more widely used than classical scaling, the latter procedure still has an important role in providing starting configurations for procedures such as nonmetric MDS.

Bibliography

Acevedo, M. and C. Restrepo (2008). Land-cover and land-use change and its contribution to the large-scale organization of Puerto Rico's bird assemblages. *Diversity and Distributions 14*, 114–122.

Agresti, A. (1992). Analysis of ordinal paired comparison data. *Applied Statistics 41*, 287–297.

Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on Automatic Control 19*, 716–723.

Alfonso-Sanchez, M., A. Perez-Miranda, and R. Herrera (2008). Autosomal microsatellite variability of the arrernte people of australia. *American Journal of Human Biology 20*, 91–99.

Amaratunga, D., J. Cabrera, and V. Kovtun (2008). Microarray learning with abc. *Biostatistics 9*, 128–136.

Ambrosi, K. and J. Hansohm (1987). *Operations Research Proceedings 1986*, Chapter Ein dynamischer Ansatz zur Repraesentation von Objekten. Berlin: Springer-Verlag.

Anon (2005). Sensory analysis - methodology - paired comparison test, iso 5495-2005.

Ashmore, R., R. Griffo, R. Green, and A. Moreno (2007). Dimensions and categories underlying thinking about college student types. *Journal of Applied Social Psychology 37*, 2922–2950.

Axelsson, Ö. (2007). Towards a psychology of photography: dimensions underlying aesthetic appeal of photographs. *Perceptual and Motor Skills 105*, 411–434.

Blake, C. (2008). Individual differences in the conceptualisation of food across eating contexts. *Food Quality and Preference 19*(1), 62–70.

Boeckenholt, U. (1990). Multivariate Thurstonian models. *Psychometrika 55*, 391–403.

Boeckenholt, U. and W. Dillon (1997). Modelling within-subject dependencies in ordinal paired comparison data. *Psychometrika 62*, 411–434.

Borg, I. and P. Groenen (2005). *Modern Multidimensional Scaling: Theory and Applications.* NY: Springer-Verlag.

Bove, G. (2005). Approaches to asymmetric multidimensional scaling with external information. In S. Zani, A. Cerioli, M. Riani, and M. Vichi (Eds.), *Data Analysis, Classification and the Forward Search*, pp. 69–76. Berlin: Springer-Verlag.

Bradley, R. (1955). Rank analysis of incomplete block designs. iii. some large-sample results on estimation and power for a method of paired comparisons. *Biometrika 42*, 450–470.

Bradley, R. and A. El-Helbawy (1976). Treatment contrasts in paired comparisons: basic procedures with application to factorials. *Biometrika 63*, 255–262.

Bradley, R. and M. Terry (1952). Rank analysis of incomplete block designs. i. the method of paired comparisons. *Biometrika 39*, 324–345.

Burnham, K. and D. Anderson (2004a). *Model Selection and Multi-model Inference: A Practical Information-theoretic Approach.* Springer-Verlag New York Inc.

Burnham, K. and D. Anderson (2004b). Multimodel inference: Understanding aic and bic in model selection. Technical report, Amsterdam Workshop on Model Selection.

Butler, R. and P. deMaynadier (2008). The significance of littoral and shoreline habitat integrity to the conservation of lacustrine damselflies (odonata). *Journal of Insect Conservation 12*, 23–36.

Cailliez, F. (1983). The analytical solution of the additive constant problem. *Psychometrika 48*, 305–308.

Carroll, J. and J. Chang (1970). Analysis of individual differences in multidimensional scaling via an n-way generalisation of 'Eckhart-Young' decomposition. *Psychometrika 35*, 283–319.

Causeur, D. and F. Husson (2005). A 2-dimensional extension of the Bradley-Terry model for paired comparisons. *Journal of Statistical Planning and Inference 135*, 245–259.

Chen, M.-F., G.-H. Tzeng, and C. Ding (2008). Combining fuzzy ahp with mds in identifying the preference similarity of alternatives. *Applied Soft Computing 8*, 110–117.

Cooke, T., F. Jäkel, C. Wallraven, and H. Bülthoff (2007). Multimodal sim-

ilarity and categorization of novel, three dimensional objects. *Neuropsychologia 45*, 484–495.

Cormack, R. (1971). A review of classification (with discussion). *Journal of the Royal Statistical Society, A 134*, 321–367.

Cox, T. (2005). *Introduction to Multivariate Analysis.* Hodder Arnold.

Cox, T. and M. Cox (2000). *Multidimensional Scaling.* CRC Press Inc.

Coxon, A. (1982). *The Users' Guide to Multidimensional Scaling.* London: Heinemann.

Crenshaw, H. and L. Edelstein-Keshet (1993). Orientation of helical motion ii. changing the direction of the axis of motion. *Bulletin of Mathematical Biology 55*(1), 213–230.

Cui, C. (2001). On the repeatability of paired comparison based scaling methods. *ISTs 2001 PICS Conference Proceedings*, 113–118.

Cunningham, L. (2006). Customer perceptions of service dimensions: cross-cultural analysis and perspective. *International Marketing Review 23*, 192–210.

David, H. (1988). *The method of paired comparisons* (2nd ed.). Hodder Arnold.

Davidson, R. (1970). Extending the Bradley-Terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association 65*, 317–328.

Davidson, R. and R. Beaver (1977). On extending the Bradley-Terry model to incorporate within-pair order effects. *Biometrics 33*, 693–702.

Davidson, R. and R. Bradley (1969). Multivariate paired comparisons: the extension of a univariate model and associated estimation and test procedures. *Biometrika 56*, 81–95.

Davidson, R. and R. Bradley (1971). A regression relationship for multivariate paired comparisons. *Biometrika 58*, 555–560.

DeJordy, R., S. Borgatti, C. Roussin, and D. Halgin (2007). Visualising proximity data. *Field Methods 19*(3), 239–263.

Della Bella, V., M. Bazzanti, M. Dowgiallo, and M. Iberite (2008). Macrophyte diversity and physico-chemical characteristics of tyrrhenian coast ponds in central italy: implications for conservation. *Hydrobiologia 597*, 85–95.

Dennis, J. and H. Mei (1979). Two new unconstrained optimisation algorithms which use function and gradient values. *Journal of Optimization Theory and Application 28*, 453–482.

DeSarbo, W., R. Grewal, and J. Wind (2006). Who competes with whom? a demand-based perspective for identifying and representing asymmetric competition. *Strategic Management Journal 27*(2), 101–129.

Ding, C. (2007). Modeling growth data using multidimensional scaling profile analysis. *Quality and Quantity 41*, 891–903.

Dittrich, R., B. Francis, R. Hatzinger, and W. Katzenbeisser (2006). Modelling dependency in multivariate paired comparisons: A log-linear approach. *Mathematical Social Sciences 52*, 197–209.

Dittrich, R., R. Hatzinger, and W. Katzenbeisser (1998). Modelling the

effect of subject-specific covariates in paired comparison studies with an application to university ranking. *Applied Statistics 47*, 511–525.

Dixon, L., C. Hamilton-Giachritsis, and K. Browne (2008). Classifying partner femicide. *Journal of Interpersonal Violence 23*, 74–93.

Dykstra, O., J. (1960). Rank analysis of incomplete block designs: a method of paired comparisons employing unequal repetitions in pairs. *Biometrics 16*, 176–188.

Eckart, C. and G. Young (1936). Approximation of one matrix by another of lower rank. *Psychometrika 1*, 211–218.

Everitt, B. and G. Dunn (2001). *Applied Multivariate Data Analysis.* Hodder Arnold.

Everitt, B. and S. Rabe-Hesketh (1997). *The Analysis of Proximity Data.* Hodder Arnold.

Faye, P., D. Bremaud, E. Teillet, P. Courcoux, A. Giboreau, and H. Nicod (2006). An alternative to external preference mapping based on consumer perceptive mapping. *Food Quality and Preference 17*, 604–614.

Fienberg, S. (1979). Log linear representation for paired comparison models with ties and within-pair order effects. *Biometrics 35*, 479–481.

Fletcher, R. (1987). *Practical Methods of Optimization.* John Wiley and Sons, Chichester.

Ford, Jr, L. (1957). Solution of a ranking problem from binary comparisons. *American Mathematical Monthly 64*, 28–33.

Forster, M. (2000). Key concepts in model selection: performance and generalizability. *Journal of Mathematical Psychology 44*, 205–231.

Francis, B., R. Dittrich, R. Hatzinger, and R. Penn (2002). Analysing partial ranks by using smoothed paired comparison methods: an investigation of value orientation in europe. *Applied Statistics 51*, 319–336.

Friendly, M. (2007). He plots for multivariate linear models. *Journal of Computational and Graphical Statistics 16*(2), 421–444.

Gabriel, K. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika 58*, 453–467.

Gacula, M., S. Rutenbeck, P. Lana, A. Resurreccion, and H. Moskowitz (2007). The just-about-right intensity scale: Functional analyses and relation to hedonics. *Journal of Sensory Studies 22*(2), 194–211.

Gay, D. (1983). Subroutines for unconstrained minimization. *ACM Transactions on Mathematical Software 9*, 503–524.

Giragame, C., C. Marasinghe, A. Madurapperuma, D. Wanasinghe, S. Herath, and O. Minetada (2006). Cross-language similarity between perceptual and semantic structures of color tones. *2006 IEEE International Conference on Systems, Man And Cybernetics, Vols 1-6 Proceedings*, 345–351.

Gower, J. (1971). A general coefficient of similarity and some of its properties. *Biometrics 27*, 857–874.

Gower, J. (1985). Properties of euclidean and non-euclidean distance matrices. *Linear Algebra and Its Applications 67*, 81–97.

Gower, J. and G. Dijksterhuis (2004). *Procrustes Problems.* Oxford University Press.

Gower, J. and D. Hand (1996). *Biplots.* London: Chapman And Hall.

Green, B. (1952). The orthogonal approximation of an oblique structure in factor analysis. *Psychometrika 17*, 429–449.

Grönen, P. and W. Heiser (1996). The tunneling method for global optimization in multidimensional scaling. *Psychometrika 61*, 529–550.

Guggenheimer, H. (1977). *Differential Geometry.* Dover.

Hjorth, U. (1982). Model selection and forward validation. *Scandinavian Journal of Statistics 9*(2), 95–105.

Hook, P. (2007, June). Visualizing the topic space of the united states supreme court. In *Proceedings of the ISSI 2007: 11th International Conference of the International Society for Scientometrics and Informetrics, Vols I and II*, pp. 387–396.

Hosobe, H. (2007). Analysis of a high-dimensional approach to interactive graph drawing. *Asia-Pacific Symposium on Visualisation 2007, Proceedings*, 93–96.

Hunter, D. (2004). Mm algorithms for generalised Bradley-Terry models. *Ann.Statist. 32*, 386–408.

Hymann, H. and H. Lawless (1999). *Sensory Evaluation in Food: Principles and Practices.* Aspen Publishers Inc, U.S.

Imray, P., W. Johnson, and G. Koch (1976). Incomplete contingency approach to paired comparison experiments. *Journal of American Statistical Association 71*, 614–623.

Iyer, B. and C. Vishveshwara (1993). Frenet-Serret description of gyroscopic precession. *Physical Review 48*, 5706–5720.

Jobson, J. (1992). *Applied multivariate data analysis Volume II: Categorical and multivariate methods.* Springer-Verlag, NY.

Kagie, M., M. van Wezel, and P. Groenen (2007, September). Online shopping using a two dimensional product map. *E-Commerce and Web Techologies, Proceedings 4655*, 89–98.

Kar, P., P. Srivastave, A. Awasthi, and S. Urs (2008). Genetic variability and association of ISSR markers with some biochemical traits in mulberry (Morus spp.) genetic resources available in India. *Tree Genetics and Genomes 4*, 75–83.

Kendall, D. (1989). A survey of the statistical theory of shape. *Statistical Science 4*(2), 87–99.

Kenkel, N. (2006). On selecting an appropriate multivariate analysis. *Canadian Journal of Plant Science 86*, 663–666.

Kim, D., W. Kim, and J. Hans (2007). A perceptual mapping of online travel agencies and preference attributes. *Tourism Management 28*(2), 591–603.

Kim, S.-K., M. Davison, and C. Frisby (2007). Confirmatory factor analysis and profile analysis via multidimensional scaling. *Multivariate Behavioral Research 42*, 1–32.

Kousgaard, N. (1976). Models for paired comparisons with ties. *Scandinavian Journal of Statistics 3*, 1–11.

Kousgaard, N. (1984). Analysis of a sound field experiment by a model of paired comparisons with explanatory variables. *Scandinavian Journal of Statistics 11*, 243–255.

Kruskal, J. (1964a). Multidimensional scaling by optimising goodness of fit to a non-metric hypothesis. *Psychometrika 29*, 1–27.

Kruskal, J. (1964b). Nonmetric multidimensional scaling: a numerical method. *Psychometrika 29*, 115–129.

Kruskal, J. and M. Wish (1978). *Multidimensional Scaling.* Sage Publications.

Krzanowski, W. (1988). *Principles of Multivariate Analysis: A User's Perspective.* Oxford Science Publications.

Lee, D. and Y. Nakamura (2007). Mimesis scheme using a monocular vision system on a humanoid robot. *Proceedings of the 2007 IEEE International Conference on Robotics and Automation, Vols 1-10*, 2162–2168.

Lespinats, S., M. Verleysen, A. Giron, and B. Fertil (2007). Dd-hds: A method for visualization and exploration of high-dimensional data. *IEEE Transactions on Neural Networks 18*, 1265–1279.

Lilensten, J., T. de Wit, P. Amblard, J. Aboudarham, F. Auchere, and M. Kretzschmar (2007). Recommendation for a set of solar euv lines to be monitored for aeronomy applications. *Annales Geophysicae 25*, 1299–1319.

Lim, J. and B. Green (2007). The psychophysical relationship between bitter taste and burning sensation: Evidence of qualitative similarity. *Chemical Senses 32*(1), 31–39.

Lin, R., C. Lin, and J. Wong (1996). An application of multidimensional scaling in product semantics. *International Journal of Industrial Ergonomics 18*, 193–204.

Liu, D., Z. Ou, G. Wang, S. Hua, and T. Su (2007). Face recognition using hierarchical isomap. *2007 IEEE Workshop on Automatic Identification Advanced Technologies, Proceedings*, 103–106.

Luce, R. (1959). *Individual choice behaviours: a theoretical analysis.* NY: J.Wiley.

Luthra, A., A. Jha, G. Ananthasuresh, and S. Vishveswara (2007). A method for computing the inter-residue interaction potentials for reduced amino acid alphabet. *Journal of Bioscience 32*, 883–889.

MacKay, D. and M. O'Mahony (2002). Sensory profiling with probabilistic multidimensional scaling. *Journal of Sensory Studies 17*, 461–481.

Marden, J. (1996). *Analysing and modeling rank data.* Chapman and Hall.

Martins, Y. and P. Pliner (2006). 'ugh! that's disgusting!': Identification of the characteristics of foods underlying rejections based on disgust. *Appetite 46*, 75–85.

Matheus, J., A. Dourado, J. Henriques, M. António, and D. Nogueira (2006). Iterative multidimensional scaling for industrial process monitoring. *2006 IEEE International Conference on Systems, Man And Cybernetics, Vols 1-6 Proceedings*, 62–67.

Matthews, J. and K. Morris (1995). An application of bradley-terry-type models to the measurement of pain. *Applied Statistics 44*, 243–255.

McBride, R., A. Watson, and B. Cox (1984). The paired-comparison method as a simple difference test. *Journal of Food Quality 6*, 285–290.

McQuarrie, A. and C. Tsai (1998). *Regression and the Time Series Model Selection.* World Scientific Publishing.

Mead, A. (1992). Review of the development of multidimensional scaling methods. *The Statistician 41*, 27–39.

Mosier, C. (1939). Determining a simlpe structure when loadings for certain tests are known. *Psychometrika 4*(2), 149–162.

Mugavin, M. (2008). Multidimensional scaling - a brief overview. *Nursing Research 57*, 64–68.

Nabi, R. (2007). Determining dimensions of reality: a concept mapping of the reality tv landscape. *Journal of Broadcasting and Electronic Media 51*, 371–390.

Napolitano, F., G. Ralconi, R. Tagliaferri, A. Ciaramella, A. Staiano, and G. Miele (2008). Clustering and visualization approaches for human cell cycle gene expression data analysis. *International Journal of Approximate Reasoning 47*, 70–84.

Nelder, J. and R. Mead (1965). A simplex method for function minimization. *Computer Journal 7*, 308–313.

Oh, M.-S. and A. Raftery (2007). Model-based clustering with dissimilarities: a Bayesian approach. *Journal of Computational and Graphical Statistics 16*, 559–585.

Park, M., J. Lee, J. Leec, and S. Song (2008). Several biplot methods applied to gene expression data. *Journal of Statistical Planning and Inference 138*(2), 500–515.

Petiot, J. and S. Grognet (2006). Product design: a vectors field approach for preference modelling. *Journal of Engineering Design 17*, 217–233.

Pittelkow, Y. and S. Wilson (2007). Visualisation of "high p, small n" data. *Computational Statistics 22*(4), 533–541.

Powell, M. (1992). A direct search optimization method that models the objective and constraint functions by linear interpolation. In *DAMTP.NA5*. Cambridge, England.

Rao, P. and L. Kupper (1967). Ties in paired-comparison experiments: a generalisation of the Bradley-Terry model. *Jounral of the American Statistical Association 62*, 194–204.

Sakamoto, Y., M. Ishiguro, and G. Kitagama (1989). Akaike information criterion statistics. *Technometrics 31*, 270–271.

Schiffman, S., M. Reynolds, and F. Young (1981). *Intorduction to Multidimensional Scaling: Theory, Methods and Applicaions*. NY: Academic Press.

Schönemann, P. and R. Carroll (1970). Fitting one matrix to another under choice of a central diilation and a rigid motion. *Psychometrika 35*, 349–366.

Sen, P. and H. David (1968). Paired comparisons for paired characteristics. *Annals of Mathematical Statistics 39*, 200–208.

Sen, P. and M. Puri (1967). On the theory of rank order tests for location in the multivariate one sample problem. *Annals of Mathematical Statistics 38*, 1216–1228.

Shepard, R. (1962a). The analysis of proximities: multidimensional scaling with an unknown distance function. 1. *Psychometrika 27*, 125–140.

Shepard, R. (1962b). The analysis of proximities: multidimensional scaling with an unknown distance function 2. *Psychometrika 27*, 219–246.

Snijders, T., M. Dormaar, W. van Schuur, C. Dijkman-Caes, and G. Driessen (1990). Distribution of some similarity coefficients for dyadic binary data in the case of assoicated attributes. *Journal of Classification 7*, 5–31.

Spivak, M. (1999). *A Comprehensive Introduction to Differential Geometry*, Volume 2. Publish or Perish, Inc.

Struik, D. (1961). *Lectures on Classical Differential Geometry.* Addison-Wesley, Reading, Mass.

Thai, B., C. Burridge, and C. Austin (2007). Genetic diversity of common carp (Cyprinus carpio L.) in Vietnam using four microsatellite loci. *Aquaculture 269*, 174–186.

Thurstone, L. (1927). Three psychophysical laws. *Pyschological Review 34*, 424–432.

Torgerson, W. (1952). Multidimensional scaling: 1. theory and method. *Psychometrika 17*, 401–419.

Torgerson, W. (1958). *Theory and Method of Scaling.* New York: Wiley.

Tsumoto, S. and S. Hirano (2007). Visualization of similarities and dissimilarities between rules using multidimensional scaling. *Knowledge-based Intelligent Information and Engineering Systems: KES 2007 - WIRN 2007, PT II, Proceedings 4693*, 978–986.

Udina, F. (2005). Interactive biplot construction. *Journal of Statistical Software 13*(5), 1–16.

Verheyen, S., E. Ameel, and G. Storms (2007). Determining the dimensionality in spatial representations of semantic concepts. *Behavior Research Methods 39*(3), 427–438.

Wang, H., B. Ge, V. Mair, D. Cal, C. Xie, Q. Zhang, H. Zhou, and H. Zhu (2007). Molecular genetic analysis of remains from Lamadong Cemetery, Liaoning, China. *American Journal of Physical Anthropology 134*(3), 404–411.

Wright, J., S. Dworjanyn, C. Rogers, P. Steinberg, J. Williamson, and A. Poore (2005). Density-dependent sea urching grazing: differential removal of species, changes in community composition and alternative community states. *Marine Ecology - Progress Series 298*, 143–156.

www.theaa.com (2008).

Yang, K.-L. and F.-L. Lin (2008). A model of reading comprehension of geometry proof. *Educational Studies in Mathematics 67*(1), 59–76.

Yoshioka, T., S. Bensmala, J. Craig, and S. Hsiao (2007). Texture perception through direct and indirect touch: An analysis of perceptual space for tactile textures in two modes of exploration. *Somatosensory and Motor Research 24*(1-2), 53–70.

Young, F. and R. Hamer (1987). *Multidimensional Scaling: History, Theory and Applications.* Hillsdale, NJ: Lawrence Erlbaum.

Young, G. and A. Householder (1938). Discussion of a set of points in terms of their mutual distances. *Psychometrika 3*, 19–22.

Zheng, X., J. Lin, S. Zapf, and C. Knapheide (2007). Visualizing user experience through "perceptual maps": Concurrent assessment of perceived usability and subjective appearance in car infotainment systems. *Digital Human Modelling 4561*, 536–545.

Zhou, R., L. An, X. Wang, W. Shao, G. Lin, W. Yu, L. Yi, S. Xu, J. Xu, and X. Xie (2007). Testing the hypothesis of an ancient Roman soldier origin of the Liqian people in northwest China: a Y-chromosome perspective. *Journal of Human Genetics 52*, 584–591.