Multidimensional Scaling

III.3

Michael A.A. Cox, Trevor F. Cox

Suppose dissimilarity data have been collected on a set of n objects or individuals, where there is a value of dissimilarity measured for each pair. The dissimilarity measure used might be a subjective judgement made by a judge, where for example a teacher subjectively scores the strength of friendship between pairs of pupils in her class, or, as an alternative, more objective, measure, she might count the number of contacts made in a day between each pair of pupils. In other situations the dissimilarity measure might be based on a data matrix. The general aim of multidimensional scaling is to find a configuration of points in a space, usually Euclidean, where each point represents one of the objects or individuals, and the distances between pairs of points in the configuration match as well as possible the original dissimilarities between the pairs of objects or individuals. Such configurations can be found using metric and non-metric scaling, which are covered in Sects. 2 and 3. A number of other techniques are covered by the umbrella title of multidimensional scaling (MDS), and here the techniques of Procrustes analysis, unidimensional scaling, individual differences scaling, correspondence analysis and reciprocal averaging are briefly introduced and illustrated with pertinent data sets.

Much of the initial impetus and theory of MDS was developed by mathematical psychologists who published many of their findings in the journal Psychometrika. Although its roots are in the behavioural sciences, MDS has now become more widely popular and has been employed in a wide variety of areas of application. This popularity is reflected by its inclusion in many computer-based statistical packages. Books on the subject include those by Borg and Groenen (1997), Cox and Cox (2001) and Young (1987).

3.1 Proximity Data

Proximity means nearness in whatever space is under consideration. The "nearness" of objects, individuals or stimuli needs defining prior to any analysis. In some situations, such as with simple Euclidean distance, this is straightforward. There are two types of basic measure of proximity, similarity and dissimilarity, with these being employed to indicate how similar or dissimilar objects are. The similarity/dissimilarity measured between two objects is a real function, resulting in similarity s_{rs} or dissimilarity δ_{rs} between the rth and sth objects. Usually all measures are taken to be non-negative. The dissimilarity of an object with itself is taken to be zero, while the similarity of an object with itself is the maximum similarity possible, with similarities usually scaled so that the maximum similarity is unity. The choice of proximity measure will depend on the problem under consideration. Sometimes the measure between two individuals is not based on any underlying observations and is totally subjective as with the teacher scoring friendship between pupils. In other situations, similarities (dissimilarities) are constructed from a data matrix for the objects. These are then called similarity (dissimilarity) coefficients. Several authors, for example Cormack (1971), Jardine and Sibson (1971), Anderberg (1973), Sneath and Sokal (1973), Diday and Simon (1976), Mardia et al. (1979), Gordon (1999), Hubalek (1982), Gower

(1985), Gower and Legendre (1986), Digby and Kempton (1987), Jackson et al. (1989), Baulieu (1989) and Snijders et al. (1990), discuss various similarity and dissimilarity measures together with their associated problems. Table 3.1 lists the more popular dissimilarities for quantitative data, where $\mathcal{X} = [x_{ri}]$ denotes the data matrix obtained for n objects on p variables $(r = 1, \ldots, n; i = 1, \ldots, p)$. The vector for the rth object is denoted by $(x)_r$ and so $\mathcal{X} = [\boldsymbol{x}_r^T]$. The $\{w_i\}$ are weights, and these and the parameter λ are chosen as required.

When all the variables are binary, it is customary to construct a similarity coefficient and then to transform this into a dissimilarity coefficient with a transformation such as $\delta_{rs} = 1 - s_{rs}$. The measure of similarity between objects r and s is based on Table 3.2. The table shows the number of variables, a, out of the total p variables where both objects score "1", the number of variables, b, where r scores "1" and s scores "0", etc. Table 3.3 gives a list of similarity coefficients based on the four counts a, b, c, d. Various situations call for particular choices of coefficients. In practice, more than one can be tried, hoping for some robustness against choice. Hubalek (1982) gives a very comprehensive list of similarity coefficients for binary data.

Table 3.1. Dissimilarity measures for quantitative data

Dissimilarity measure	Formula
Euclidean distance	$\delta_{rs} = \left\{\sum_i (x_{ri} - x_{si})^2\right\}^{1/2}$
Weighted Euclidean	$\delta_{rs} = \left\{\sum_i w_i (x_{ri} - x_{si})^2\right\}^{1/2}$
Mahalanobis distance	$\delta_{rs} = \{(\boldsymbol{x}_{ri} - \boldsymbol{x}_{si})' \Sigma^{-1} (\boldsymbol{x}_r - \boldsymbol{x}_s)\}^{1/2}$
City block metric	$\delta_{rs} = \sum_i \lvert x_{ri} - x_{si}\rvert$
Minkowski metric	$\delta_{rs} = \left\{\sum_i w_i \lvert x_{ri} - x_{si}\rvert^{\lambda}\right\}^{1/\lambda} \quad \lambda \geq 1$
Canberra metric	$\delta_{rs} = \sum_i \dfrac{\lvert x_{ri} - x_{si}\rvert}{x_{ri} + x_{si}}$
Divergence	$\delta_{rs} = \frac{1}{p} \sum_i \dfrac{(x_{ri} - x_{si})^2}{(x_{ri} + x_{si})^2}$
Bray-Curtis	$\delta_{rs} = \frac{1}{p} \dfrac{\sum_i \lvert x_{ri} - x_{si}\rvert}{\sum_i (x_{ri} + x_{si})}$
Soergel	$\delta_{rs} = \frac{1}{p} \dfrac{\sum_i \lvert x_{ri} - x_{si}\rvert}{\sum_i \max(x_{ri}, x_{si})}$
Bhattacharyya distance	$\delta_{rs} = \sqrt{\sum_i \left(\sqrt{x_{ri}} - \sqrt{x_{si}}\right)^2}$
Wave–Hedges	$\delta_{rs} = \sum_i \left(1 - \dfrac{\min(x_{ri}, x_{si})}{\max(x_{ri}, x_{si})}\right)$
Angular separation	$\delta_{rs} = 1 - \dfrac{\sum_i x_{ri} x_{si}}{\left[\sum_i x_{ri}^2 \sum_i x_{si}^2\right]^{1/2}}$
Correlation	$\delta_{rs} = 1 - \dfrac{\sum_i (x_{ri} - \bar{x}_r)(x_{si} - \bar{x}_s)}{\left[\sum_i (x_{ri} - \bar{x}_r)^2 \sum_i (x_{si} - \bar{x}_s)^2\right]^{1/2}}$

Table 3.2. Summary of binary totals

		Object s		
		1	0	
Object r	1	a	b	$a+b$
	0	c	d	$c+d$
		$a+c$	$b+d$	$p=a+b+c+d$

Table 3.3. Similarity coefficients for Binary Data

Coefficient	Formula
Braun, Blanque	$s_{rs} = \dfrac{a}{\max\{(a+b),(a+c)\}}$
Czekanowski, Sørensen, Dice	$s_{rs} = \dfrac{2a}{2a+b+c}$
Hamman	$s_{rs} = \dfrac{a-(b+c)+d}{a+b+c+d}$
Jaccard coefficient	$s_{rs} = \dfrac{a}{a+b+c}$
Kulczynski	$s_{rs} = \dfrac{a}{b+c}$
Kulczynski	$s_{rs} = \frac{1}{2}\left(\dfrac{a}{a+b} + \dfrac{a}{a+c}\right)$
Michael	$s_{rs} = \dfrac{4(ad-bc)}{(a+d)^2+(b+c)^2}$
Mountford	$s_{rs} = \dfrac{2a}{a(b+c)+2bc}$
Mozley, Margalef	$s_{rs} = \dfrac{a(a+b+c+d)}{(a+b)(a+c)}$
Ochiai	$s_{rs} = \dfrac{a}{\sqrt{(a+b)(a+c)}}$
Phi	$s_{rs} = \dfrac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$
Rogers, Tanimoto	$s_{rs} = \dfrac{a+d}{a+2b+2c+d}$
Russell, Rao	$s_{rs} = \dfrac{a}{a+b+c+d}$
Simple matching coefficient	$s_{rs} = \dfrac{a+d}{a+b+c+d}$
Simpson	$s_{rs} = \dfrac{a}{\min\{(a+b),(a+c)\}}$
Sokal, Sneath, Anderberg	$s_{rs} = \dfrac{a}{a+2(b+c)}$
Yule	$s_{rs} = \dfrac{ad-bc}{ad+bc}$

For categorical data, agreement scores can be used where, for example, if objects r and s share the same category, then $\delta_{rs} = 0$ and $\delta_{rs} = 1$ if they do not. Other, more elaborate, agreement scores can be devised.

When data are mixed, with binary, quantitative and categorical variables, Gower (1971) suggests using a general similarity coefficient,

$$s_{rs} = \frac{\sum_{i=1}^{p} w_{rsi} s_{rsi}}{\sum_{i=1}^{p} w_{rsi}}, \tag{3.1}$$

where s_{rsi} is the similarity between the rth and sth objects based on the ith variable alone and w_{rsi} is unity if the rth and sth objects can be compared on the ith variable and zero otherwise. For quantitative variables, Gower suggests $s_{rsi} = 1 - |x_{ri} - x_{si}|/R_i$, where R_i is the range of the observations for variable i. For presence/absence data, Gower suggests $s_{rsi} = 1$ if objects r and s both score "presence," and zero otherwise, while $w_{rsi} = 0$ if objects r and s both score "absence," and unity otherwise. For nominal data Gower suggests $s_{rsi} = 1$ if objects r and s share the same categorization, and zero otherwise.

Metric MDS

3.2

Given n objects with a set of dissimilarities $\{d_{rs}\}$, one dissimilarity for each pair of objects, metric MDS attempts to find a set of points in some space where each point represents one of the objects and the distances between points $\{d_{rs}\}$ are such that

$$d_{rs} = f(\delta_{rs}), \tag{3.2}$$

where f is a continuous parametric monotonic function. The function f can either be the identity function or a function that attempts to transform the dissimilarities to a distance-like form. The first type of metric scaling described here is classical scaling, which originated in the 1930s when Young and Householder (1938) showed that, starting with a matrix of distances between all pairs of the points in a Euclidean space, coordinates for the points could be found such that distances are preserved. Torgerson (1952) brought the subject to popularity using the technique for scaling, where distances are replaced by dissimilarities.

The algorithm for recovering coordinates from distances between pairs of points is as follows:

1. Form matrix $\mathcal{A} = [-\frac{1}{2}\delta_{rs}^2]$.
2. Form matrix $\mathcal{B} = \mathcal{H}\mathcal{A}\mathcal{H}$, where $\mathcal{H}$ is the centring matrix $\mathcal{H} = \mathcal{I} - n^{-1}\mathbf{1}_n\mathbf{1}_n{}^T$, with $\mathbf{1}_n$ a vector of ones.
3. Find the spectral decomposition of $\mathcal{B}$, $\mathcal{B} = \mathcal{V}\Lambda\mathcal{V}^T$, where Λ is the diagonal matrix formed from the eigenvalues of $\mathcal{B}$, and $\mathcal{V}$ is the matrix of corresponding eigenvectors.

4. If the points were originally in a p-dimensional space, the first p eigenvalues of $\mathcal{B}$ are nonzero and the remaining $n - p$ are zero. Discard these from Λ (rename as Λ_1), and discard the corresponding eigenvalues from $\mathcal{V}$ (rename as $\mathcal{V}_1$).
5. Find $\mathcal{X} = \mathcal{V}_1 \Lambda_1^{1/2}$, and then the coordinates of the points are given by the rows of $\mathcal{X}$.

As an example, rather than use distances between cities and towns in the UK, the cost of rail travel between all pairs of the following mainland terminus rail stations were used: Aberdeen, Birmingham, Blackpool, Brighton, Dover (Priory), Edinburgh, Inverness, Liverpool, London, Newcastle upon Tyne, Norwich, Plymouth, Sheffield, Southampton, Swansea. Figure 3.1 shows a plot of the stations having obtained the coordinates using the above algorithm. This solution is not unique since any translation, rotation or reflection of the configuration of points will give rise to another solution.

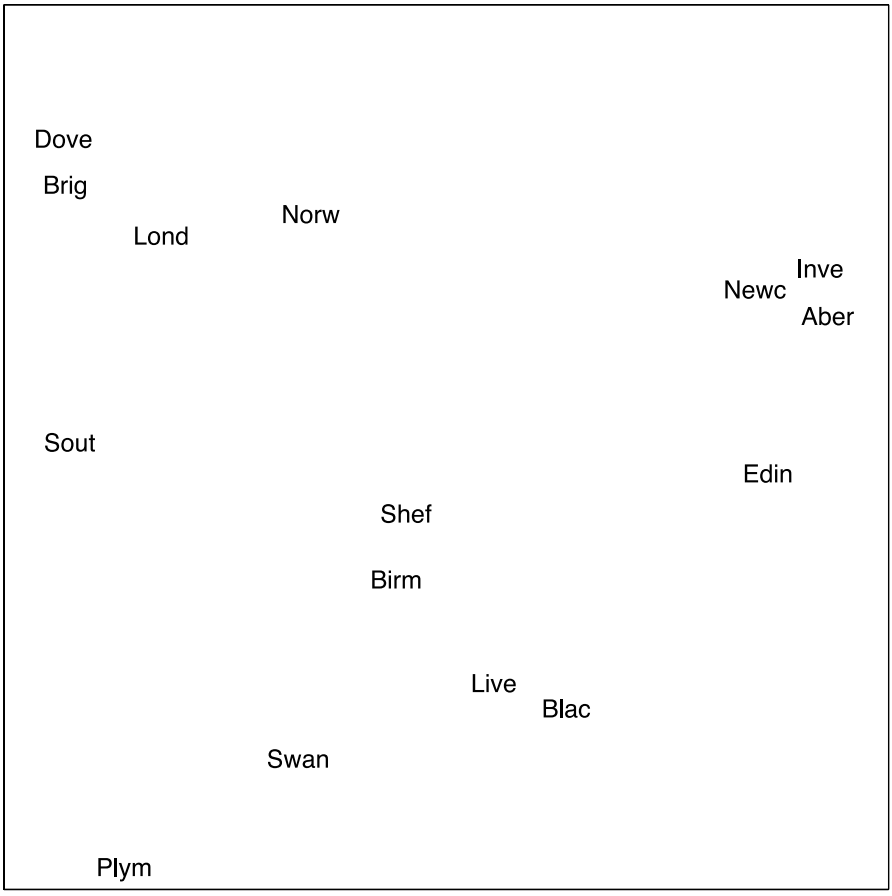

Figure 3.1. A map of rail stations using classical scaling

The plot produces a good representation of the map of the UK. The vertical axis represents West–East, while the horizontal axis runs South–North. It would appear that Newcastle upon Tyne has relocated to Scotland!

Had the exact distances between the rail stations been used in the above (and assuming the UK is in a 2-D Euclidean space!), coordinates would have been found for the stations that would have exactly reproduced the pairwise distances between them. All eigenvalues of $\mathcal{B}$ would have been zero except for the first two. In general, rather than using distances or pseudo-distances between points, classical scaling uses dissimilarities calculated between pairs of objects in place of these distances. The configuration of points obtained in a 2-D space will not usually reproduce the pairwise dissimilarities exactly, but will only approximate them. This implies that nearly all of the eigenvalues of $\mathcal{B}$ are likely to be nonzero, and some might be negative, which will occur if the dissimilarity measure is not a metric. In practice the largest two (positive) eigenvalues and their associated eigenvectors are used for the coordinates of the points. If a 3-D representation is required, then the three largest eigenvalues are used, and so on. A measure of how well the obtained configuration represents the set of pairwise dissimilarities is given by

$$\frac{\sum_{i=1}^{p} \lambda_i}{\sum_{i=1}^{n-1} |\lambda_i|} \qquad \text{or} \qquad \frac{\sum_{i=1}^{p} \lambda_i}{\sum(\text{positive eigenvalues})} \,. \tag{3.3}$$

Incidentally, if the dissimilarities are calculated as Euclidean distances, then classical scaling can be shown to be equivalent to principal component analysis.

The next example consists of 61 viruses with rod-shaped particles affecting various crops (tobacco, tomato, cucumber and others) recently employed by Ripley (1996) and originally described by Fauquet et al. (1988) and analysed by Eslava-Gomez (1989). There are 18 measurements on each virus, the number of amino acid residues per molecule of coat protein. The whole data set consists of four groups of viruses, Hordeviruses (3), Tobraviruses (6), Tobamoviruses (39) and Furoviruses (13). For brevity the initial four letters of their names will denote the four virus groups. Figure 3.2 shows a classical scaling of the data.

While Tobr and Hord form clear clusters, Furo splits into three clear groups, one of which is similar to Tobr. Similarly Toba forms two groups, one of which is similar to Tobr. The first two eigenvalues have values 6912 and 1956. The sum of all 18 significant eigenvalues is 13 597 out of a potential of 61 values. The first two dimensions correspond to 65% and hence provide a reasonable description of the data.

Another metric scaling approach is to minimize a loss function. For a Sammon map (Sammon 1969), a particular configuration of points with pairwise distances, $\{d_{rs}\}$, representing the dissimilarities $\{\delta_{rs}\}$, has loss function

$$S = \sum_{r<s} \delta_{rs}^{-1} (d_{rs} - \delta_{rs})^2 / \sum_{r<s} \delta_{rs} \,. \tag{3.4}$$

A configuration is found that has minimum loss using an appropriate optimization method such as a steepest descent method. Other loss functions have also been suggested and used. Figure 3.3 shows a Sammon map for the virus data.

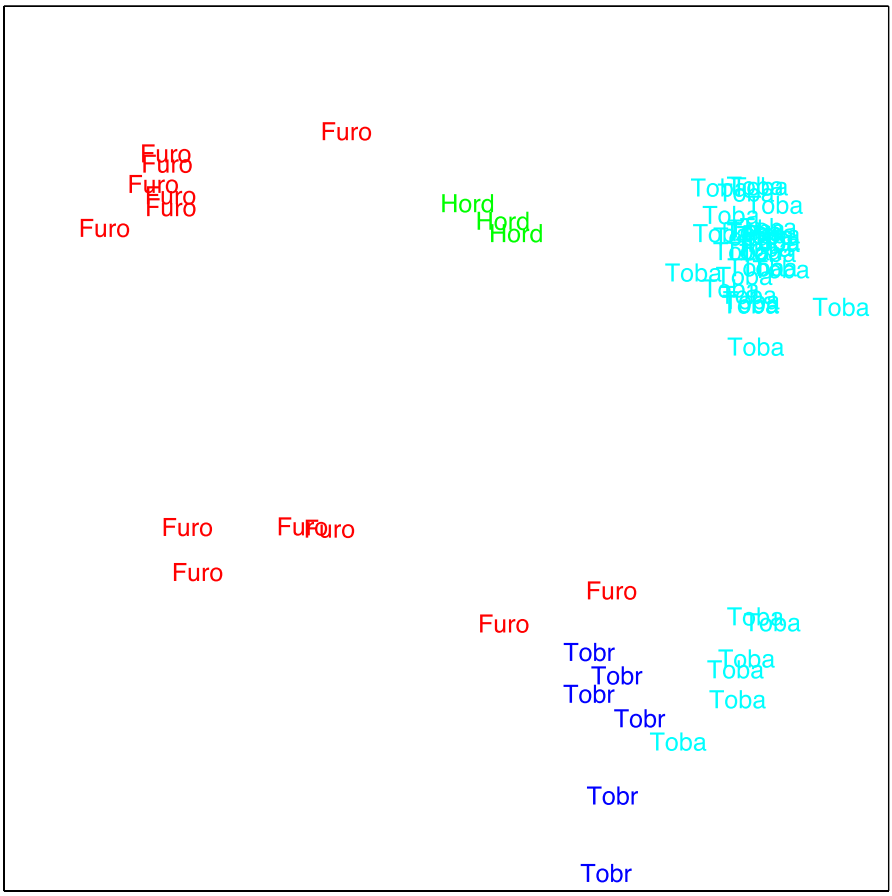

Figure 3.2. Classical scaling of virus data: Hord – Hordeviruses, Tobr – Tobraviruses, Toba – Tobaviruses, Furo – Furoviruses

While the Sammon mapping produces a structure similar to that obtained using classical scaling, the clusters are less clear cut. This differing view probably arises because the mapping is an iterative procedure and is hence dependent on the initial vector selected and the number of iterations performed.

3.3 Non-metric MDS

A non-metric approach to MDS was developed by Shepard (1962a, b) and further improved by Kruskal (1964a, b). In summary, suppose there are n objects with dissimilarities $\{\delta_{rs}\}$. The procedure is to find a configuration of n points in a space, which is usually chosen to be Euclidean, so that each object is represented by a point

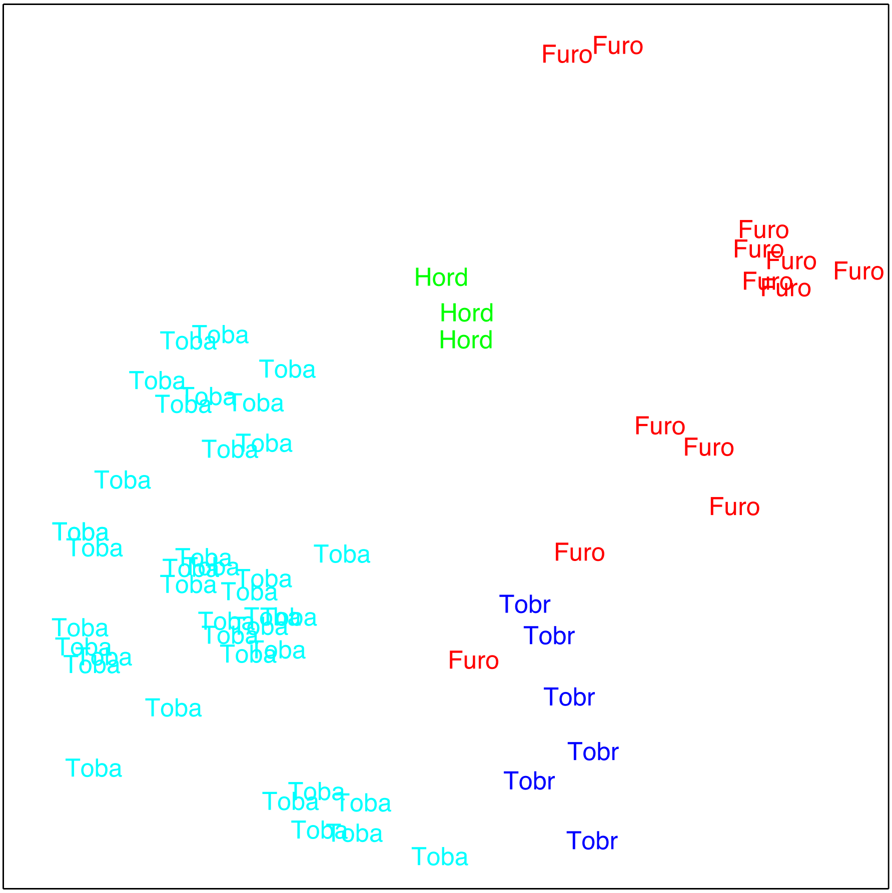

Figure 3.3. Sammon mapping of the virus data: Hord – Hordeviruses, Tobr – Tobraviruses, Toba – Tobaviruses, Furo – Furoviruses

in the space. A configuration is sought so that distances between pairs of points $\{d_{rs}\}$ in the space match "as well as possible" the original dissimilarities $\{\delta_{rs}\}$. Here matching means the rank order of $\{d_{rs}\}$ matches the rank order of $\{\delta_{rs}\}$ as best as possible. The matching of the distances $\{d_{rs}\}$ to the dissimilarities $\{\delta_{rs}\}$ for a particular configuration is measured by the STRESS (S), where

$$S = \sqrt{\frac{\sum_{r,s}(d_{rs} - \hat{d}_{rs})^2}{\sum_{rs} d_{rs}^2}}. \tag{3.5}$$

Here, $\{\hat{d}_{rs}\}$ is the primary monotone least-squares regression of $\{d_{rs}\}$ on $\{\delta_{rs}\}$, also known as isotonic regression. Details of this regression are not entered into here, but an example can be seen in Fig. 3.5, the Shepard plot. Further details can be found

in Cox and Cox (2001), Borg and Groenen (1997) and elsewhere. A configuration is found that minimizes S, usually using a gradient descent approach.

The rail data used for classical scaling were analysed using non-metric MDS. Figure 3.4 shows the configuration obtained, and again the solution is arbitrary up to translation, rotation and reflection. The STRESS associated with the optimum solution is 10 %. It should be noted that 1000 randomly selected starting points were employed to ensure that the true optimum has been obtained.

A Shepard plot may also be utilized to assess the procedure. This is simply a plot of d_{rs} and $\hat{d}_{rs}$ against δ_{rs} and is shown in Fig. 3.5 for the rail station data. It shows how well the distances within the configuration match the original dissimilarities according to rank order. It makes the monotonic least-squares regression fit particularly clear by joining the δ_{rs} and $\hat{d}_{rs}$ pairs.

The next section gives a more detailed example and shows how the quality of the fit of the model can be investigated.

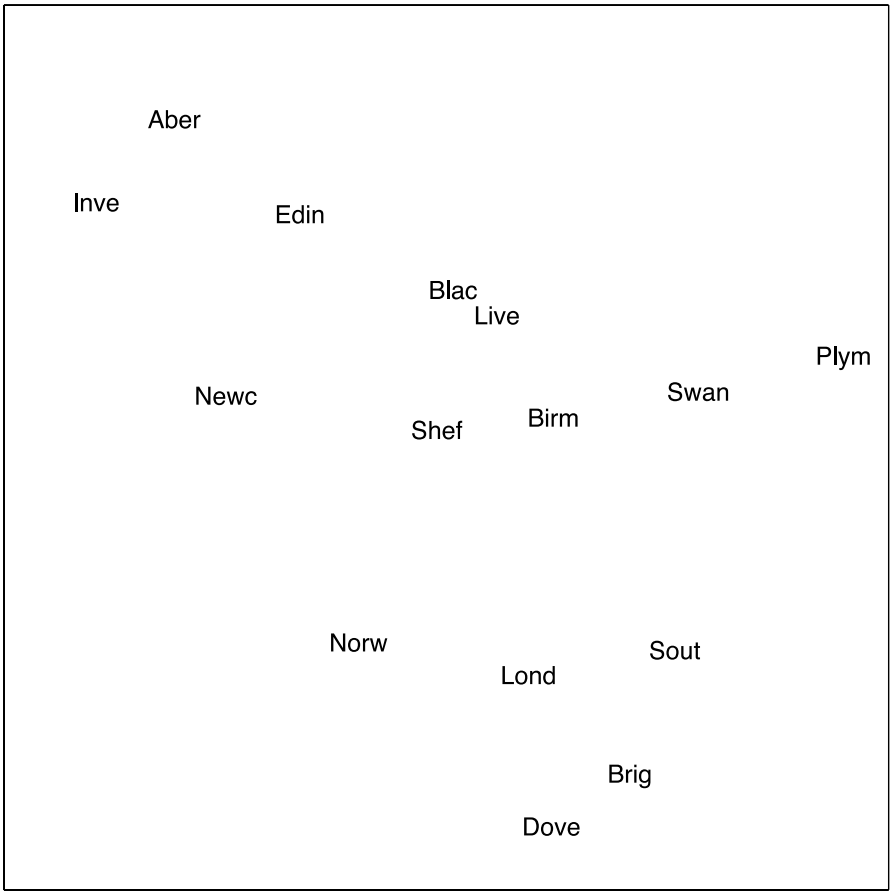

Figure 3.4. A map of rail stations from non-metric MDS

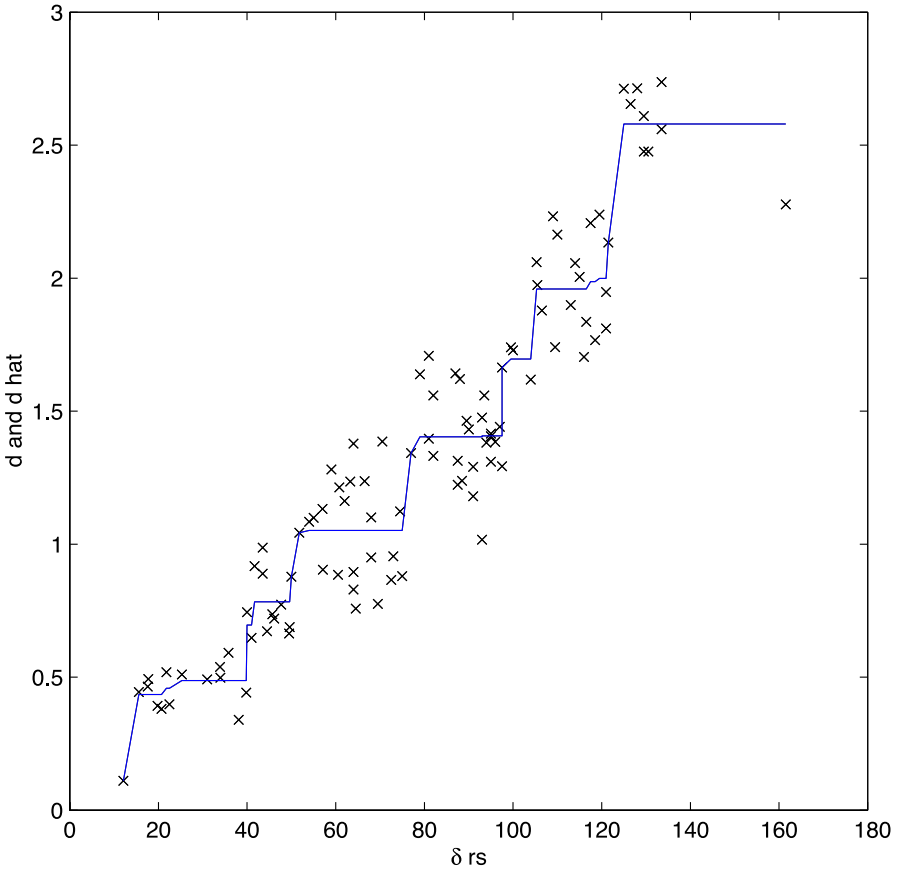

Figure 3.5. The shepard plot for the rail station data

Example: Shakespeare Keywords

3.4

Cox (2005) uses classical scaling on frequency counts of keywords from 20 Shakespeare plays. A similar analysis was carried out but using non-metric scaling with dissimilarity defined by Euclidean distance calculated from the frequencies. Figure 3.6 shows the configuration for the keywords, but where "lord" and "king" have been excluded from the analysis since their inclusion forced all other words into a single cluster producing zero STRESS. The STRESS obtained after exclusion was 11%. It is difficult to plot the configuration in such a small page area, and the only clearly visible words are related to characters plus "god" and "dead." Figure 3.7 shows the configuration obtained for plays where the role of keywords and plays has been interchanged. The STRESS for this configuration is 10%. It would appear that Hamlet is closest to the historical plays, while the tragedies and comedies are hard to separate.

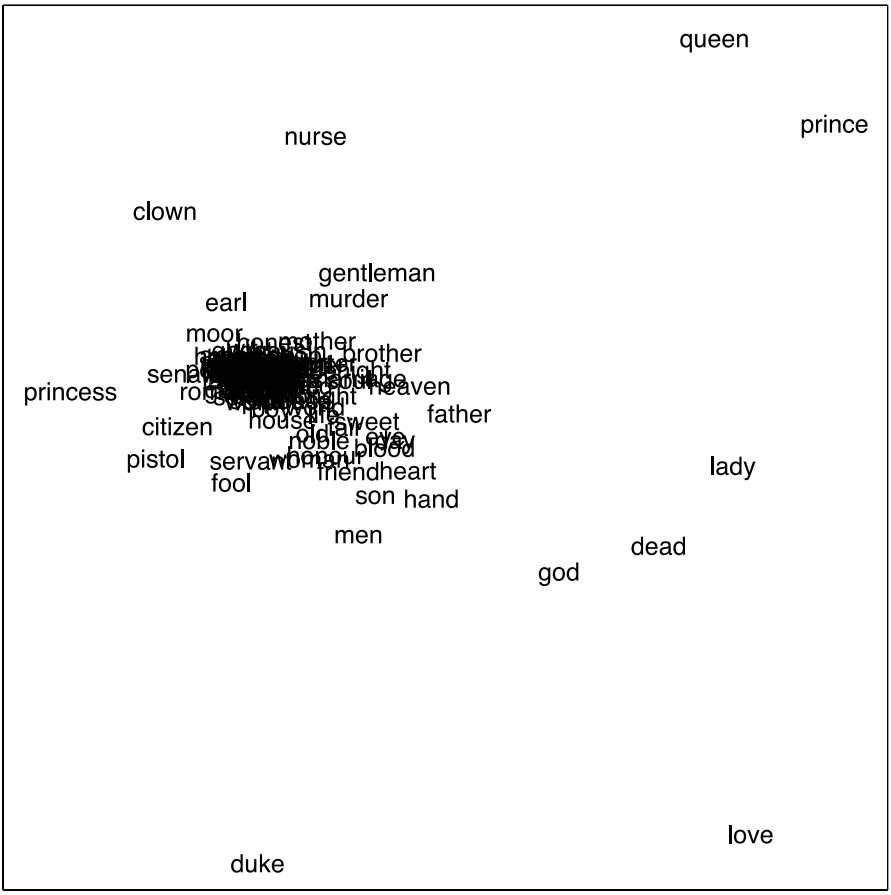

Figure 3.6. Non-metric MDS of Shakespearean keywords

In practice, the number of dimensions necessary for an informative solution is often investigated using a plot of STRESS against number of dimensions (scree plot). This is shown in Fig. 3.8 for the Shakespeare plays and indicates that three or more dimensions may have been preferable. However, then there is the eternal problem of displaying the configuration in just two dimensions.

Although not often carried out in practice, it is possible to investigate the MDS analyses further by looking at outlying or ill-fitting points. Figure 3.9 shows the Shepard plot of d_{rs} and $\hat{d}_{rs}$ against δ_{rs} for the play data. Two points appear removed from the main cluster of points showing large values of $d_{rs} - \hat{d}_{rs}$. These are (tim, h41) and (ham, h41). Figure 3.10 shows a histogram of the values of $d_{rs} - \hat{d}_{rs}$, showing a normal type distribution with a few outliers.

Table 3.4 gives the mean squared differences of $d_{rs} - \hat{d}_{rs}$ for each play (i.e. averaged over s for each r). Henry IV, Part 1 appears discordant with the other plays.

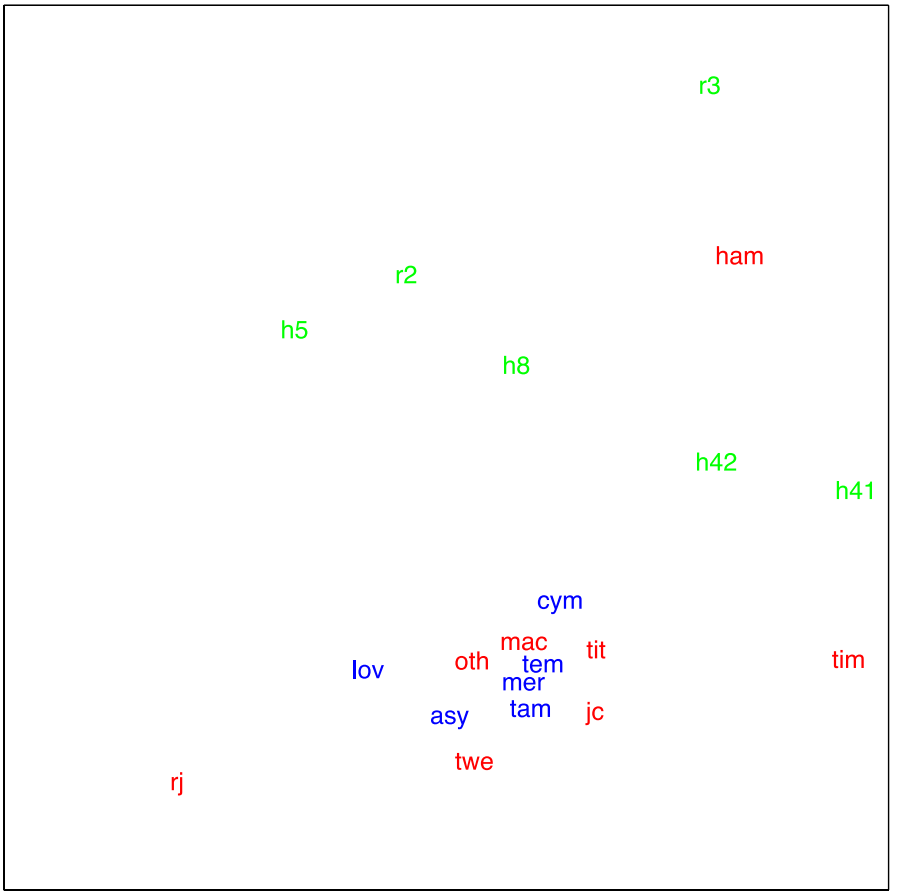

Figure 3.7. Nonmetric MDS of Shakespearean plays: (plays colored *blue* are comedies) asy – As You Like It, cym – Cymbeline, lov – Love's Labours Lost, mer – Merchant of Venice, tam – Taming of the Shrew, tem – The Tempest; (plays colored *green* are historical plays) h41 – Henry IV, Part 1, h42 – Henry IV, Part 2, h5 – Henry V, h8 – Henry VIII, r2 – Richard II, r3 – Richard III; (plays colored *red* are tragedies) ham – Hamlet, jc – Julius Caesar, mac – Macbeth, oth – Othello, rj – Romeo and Juliet, tim – Timon of Athens, tit – Titus Andronicus, twe – Twelfth Night

To find the effect of each δ_{rs} on the fitting of the configuration, each δ_{rs} can be left out of the analysis in turn and the STRESS re-calculated. Also, the resulting configuration each time can be matched to the original and the resulting value of the Procrustes analysis (Sect. 5) noted. The lowest STRESS that was obtained was 8.4 % when the dissimilarity between Timon of Athens and Henry IV, Part 1 was removed. The next lowest STRESS was 8.6 % when the dissimilarity between Hamlet and Henry IV, Part 1 was removed. For all other dissimilarities, the STRESS was back to approximately the original value. A similar exercise was carried out removing whole plays at a time. Table 3.5 shows the STRESS values obtained each time a play was removed.

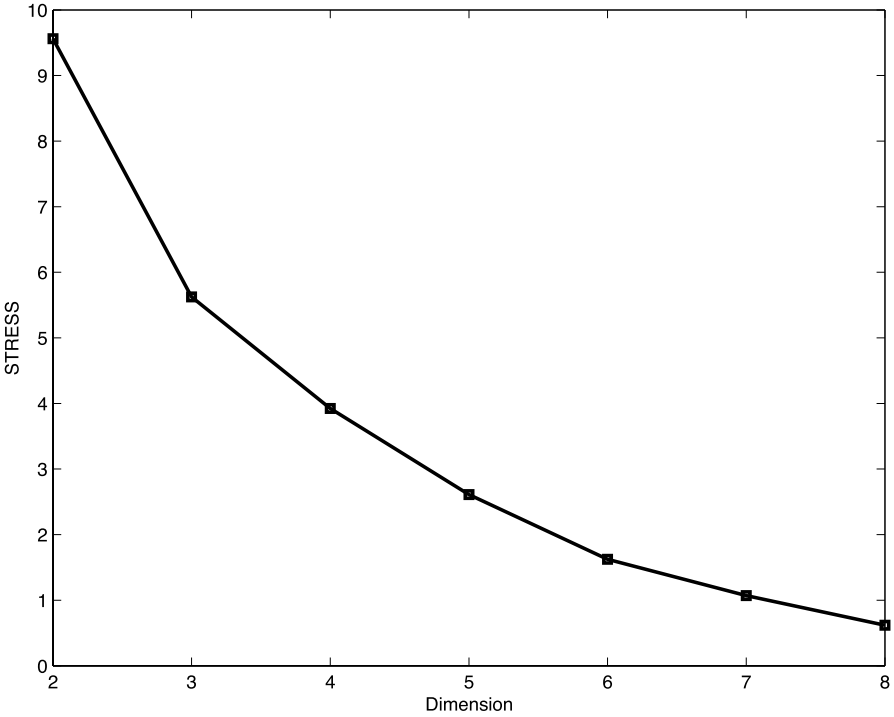

Figure 3.8. Screen plot for Shakespearean data

Table 3.4. Mean $(d_{rs} - \hat{d}_{rs})^2 \times 1000$ for each Shakespearean play

play	mean	play	mean
tam	4.5	r2	14.0
mer	5.3	r3	14.4
tem	7.2	twe	15.3
mac	7.4	lov	19.5
tit	8.8	h5	19.9
asy	9.4	cym	29.0
h8	9.7	rj	29.8
h42	9.8	ham	42.1
oth	10.7	tim	46.5
jc	11.6	h41	69.9

Again if Henry IV, Part 1 is removed, then the STRESS is reduced, showing this play is the worst fitting one. At the other end of the scale, removing Richard III increases the STRESS to 11 %, showing that that play fits the model well.

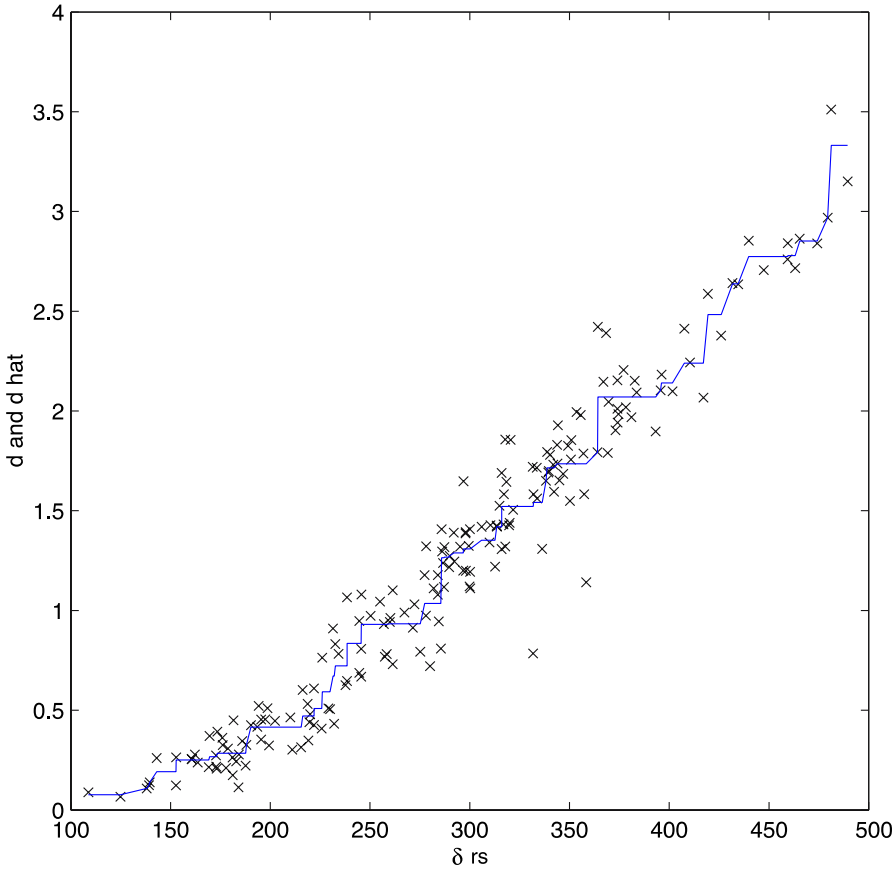

Figure 3.9. Shepard plot of Shakespearean data

Table 3.5. STRESS×100 obtained when each play is removed in turn

Play	STRESS	Play	STRESS
h41	7.24	ort	9.52
tim	8.44	h42	9.58
cym	8.61	mac	9.60
ham	8.69	h8	9.61
rj	9.16	tem	9.62
lov	9.23	asy	9.64
h5	9.30	mer	9.70
twe	9.38	r2	9.71
jc	9.48	tam	9.76
tit	9.51	r3	10.76

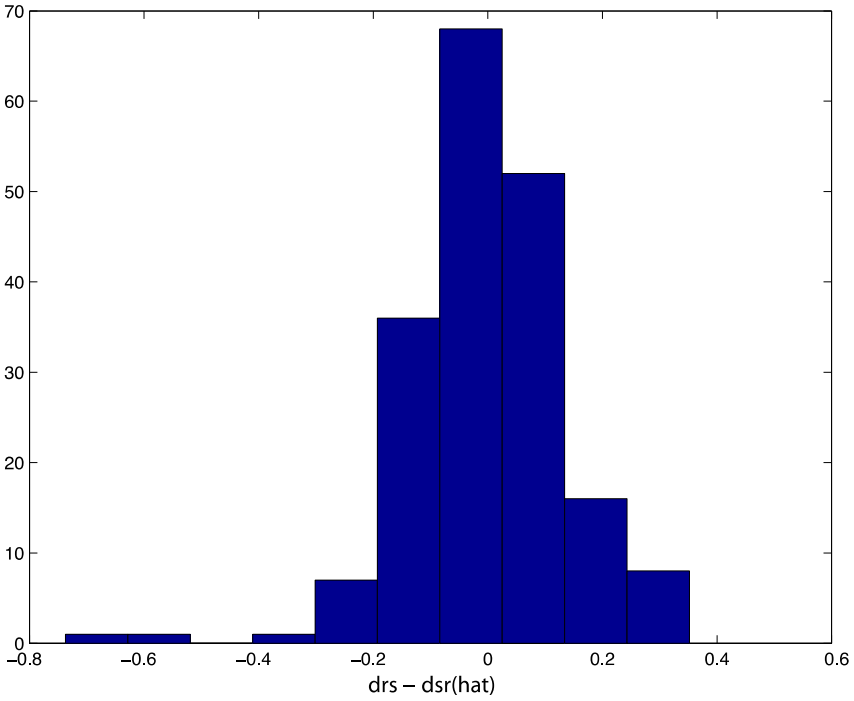

Figure 3.10. Histogram of $d_{rs} - \hat{d}_{rs}$ for Shakespearean data

3.5 Procrustes Analysis

The classical scaling analysis and the non-metric scaling of the terminal train station data produced different, but similar, configurations of points. Since arbitrary translations, rotations and reflections of these configurations give equally valid solutions. In order to make a clear visual comparison of the two, we need to match one configuration with the other. This is achieved using Procrustes analysis. Procrustes analysis finds the isotropic dilation, translation, reflection and rotation that best match one configuration to another. A detailed account of this and allied methods is given by Gower and Dijksterhuis (2004). (According to Greek mythology Procrustes was an innkeeper living near Athens who would subject his guests to extreme measures to make them fit his beds. If they were too short, he would stretch them, or if they were too long, he would cut off their legs.)

Suppose a configuration of n points in a q-dimensional Euclidean space, with coordinates given by the n by q matrix $\mathcal{X}$, is to be matched to another configuration of points in a p-dimensional Euclidean space $(p \geq q)$, with coordinates given by the n by p matrix $\mathcal{Y}$. Note, it is assumed that the rth point in the X space is in a one-to-one correspondence with the rth point in the Y space. First $p - q$ columns of zeros are added to the end of matrix $\mathcal{X}$ in order to give the matrices the same dimensions.

A measure of the discrepancy between the two configurations is given by the sum of squared distances, R^2, between corresponding points in the two spaces, i.e.

$$R^2 = \sum_{r=1}^{n} (\boldsymbol{y}_r - \boldsymbol{x}_r)^T (\boldsymbol{y}_r - \boldsymbol{x}_r)\,, \tag{3.6}$$

where $\mathcal{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^T$, $\mathcal{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n]^T$ and $\boldsymbol{x}_r$ and $\boldsymbol{y}_r$ are the coordinate vectors of the rth point in the two spaces.

The points in the X space are dilated, translated, rotated and reflected to new coordinates, $\boldsymbol{x}'$, where

$$\boldsymbol{x}'_r = \rho \mathcal{A}^T (x)_r + \boldsymbol{b}\,, \tag{3.7}$$

ρ is a dilation, $\mathcal{A}$ is an orthogonal matrix giving a rotation and possibly a reflection and $\boldsymbol{b}$ is a translation. The optimal values of these that minimizes R^2 are summarized in the following procedure:

1. (Optimum translation) Place the centroids of the two configurations at the origin.
2. (Optimum rotation) Find $\mathcal{A} = (\mathcal{X}^T \mathcal{Y} \mathcal{Y}^T \mathcal{X})^{1/2} (\mathcal{Y}^T \mathcal{X})^{-1}$ and rotate $\mathcal{X}$ to $\mathcal{X}\mathcal{A}$.
3. (Optimum scaling) Scale the $\mathcal{X}$ configuration by multiplying each coordinate by $\rho = \mathrm{tr}(\mathcal{X}^T \mathcal{Y} \mathcal{Y}^T \mathcal{X}) / \mathrm{tr}(\mathcal{X}^T \mathcal{X})$.
4. Calculate the Procrustes statistic

$$R^2 = 1 - \{\mathrm{tr}(\mathcal{X}^T \mathcal{Y} \mathcal{Y}^T \mathcal{X})^{1/2}\}^2 / \mathrm{tr}(\mathcal{X}^T \mathcal{X}) \mathrm{tr}(\mathcal{Y}^T \mathcal{Y})\,. \tag{3.8}$$

The value of R^2 can be between 0 and 1, where 0 implies a perfect matching of the configurations. The larger the value of R^2, the worse the match.

Procrustes analysis was used on the cost of rail travel data. Figure 3.11 shows the non-metric scaling result (Fig. 3.4) matched to the metric (Fig. 3.1). In this case the Procrustes statistic is 0.06, showing that the point configurations are remarkably similar.

Extensions to basic Procrustes analysis of matching one configuration to another include weighting of points and axes, the allowance of oblique axes and the matching of more than two configurations; see Cox and Cox (2001) or Gower and Dijksterhuis (2004) for a detailed account of the area.

Unidimensional Scaling

3.6

When the space in which the points representing objects or individuals has only one dimension, the scaling technique becomes that of unidimensional scaling. The loss function to be minimized is

$$S = \sum_{r<s} (\delta_{rs} - |x_r - x_s|)^2\,. \tag{3.9}$$

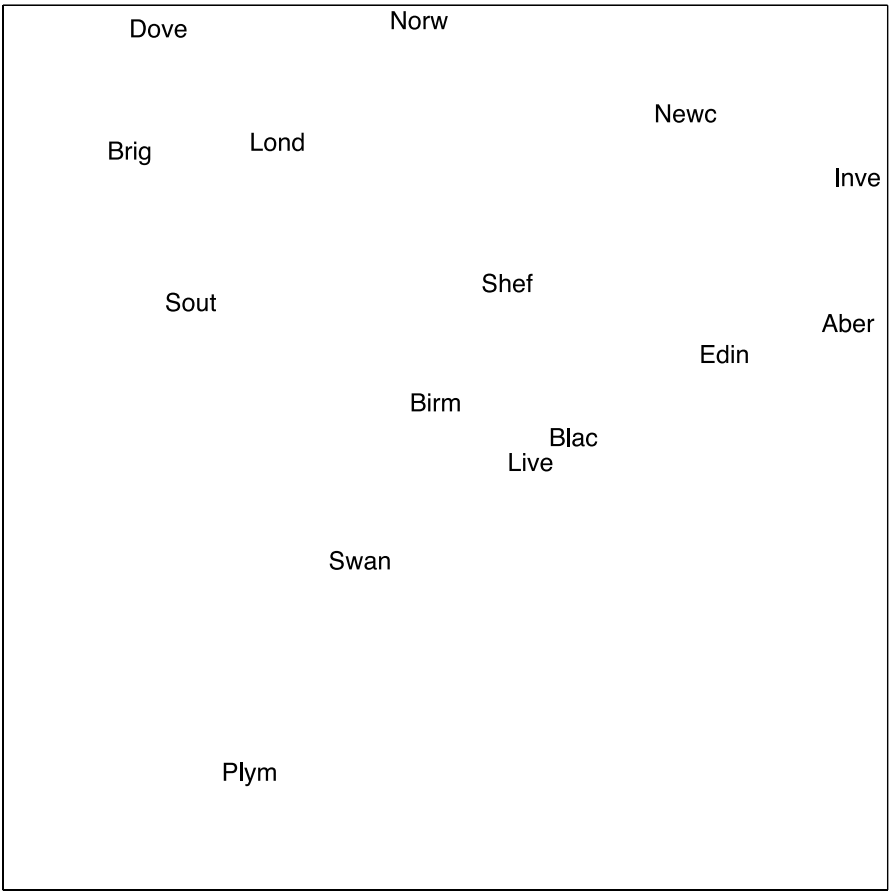

Figure 3.11. Procrustes rotation of metric onto non-metric scaling for rail cost

Minimizing S can be difficult because of the possible large number of local minima. Various algorithms have been proposed, for example Guttman (1968), Hubert and Arabie (1986, 1988) and Lau et al. (1998). Here we use the algorithm by Guttman (1968) for the following example.

The data used to illustrate this method are ratings for World War II politicians and are the lower triangle given by Everitt and Dunn (1983). Figure 3.12 shows the resulting unidimensional scaling obtained.

For clarity, the points have been plotted on a diagonal, which represents a linear axis starting from Mao Tse Tung at 6.75 and ending at Mussolini at 5.67. The countries the leaders represent have also been added. What is interesting is that Stalin is most closely identified with Hitler and Mussolini as opposed to his UK/US World War II allies.

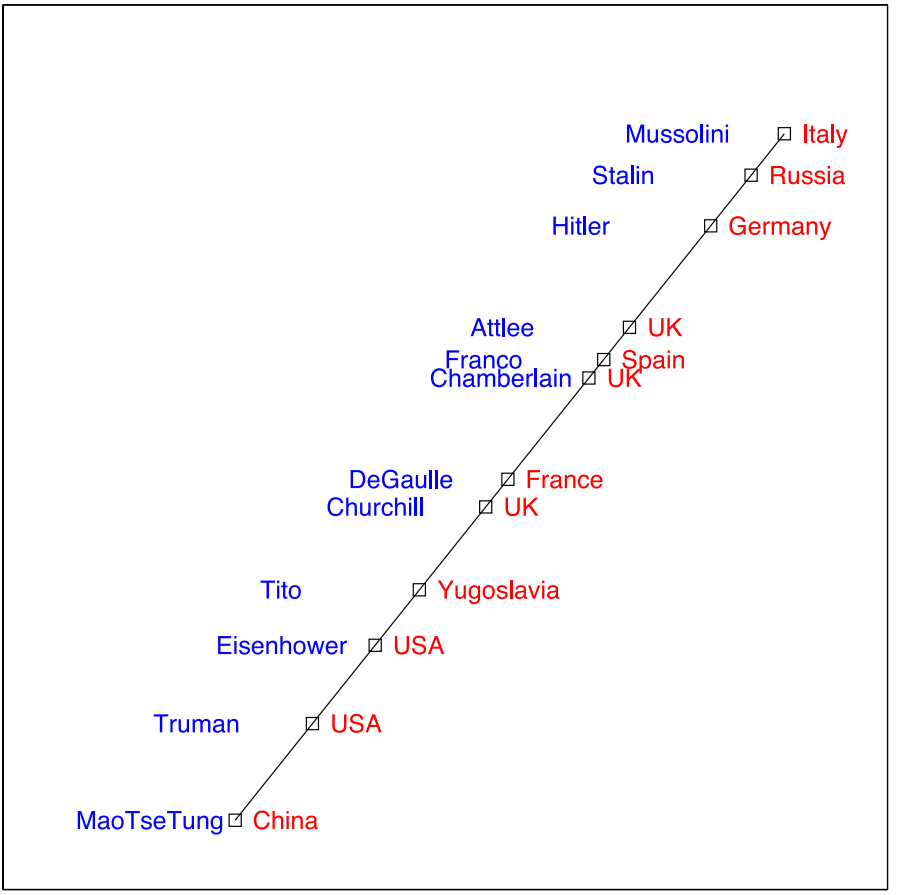

Figure 3.12. Unidimensional scaling of World War II political leaders

INDSCAL

3.7

Suppose data consist of several sets of dissimilarities between objects, the same objects in each case. For example, several panellists assess the dissimilarities between all pairs of a set of products. MDS could be applied to each panelist's data resulting in many configurations. A better approach might be to combine the dissimilarities in some manner.

Carroll and Chang (1970) proposed a metric model comprising two spaces: a group stimulus space and a subject's (or individual's) space, both chosen to be of the same dimension. Points in the group stimulus space represent the objects or stimuli and form an "underlying" configuration. The individuals are represented as points in the subject's space. The coordinates of each individual are the weights required to give the weighted Euclidean distances between the points in the stimulus space, the values

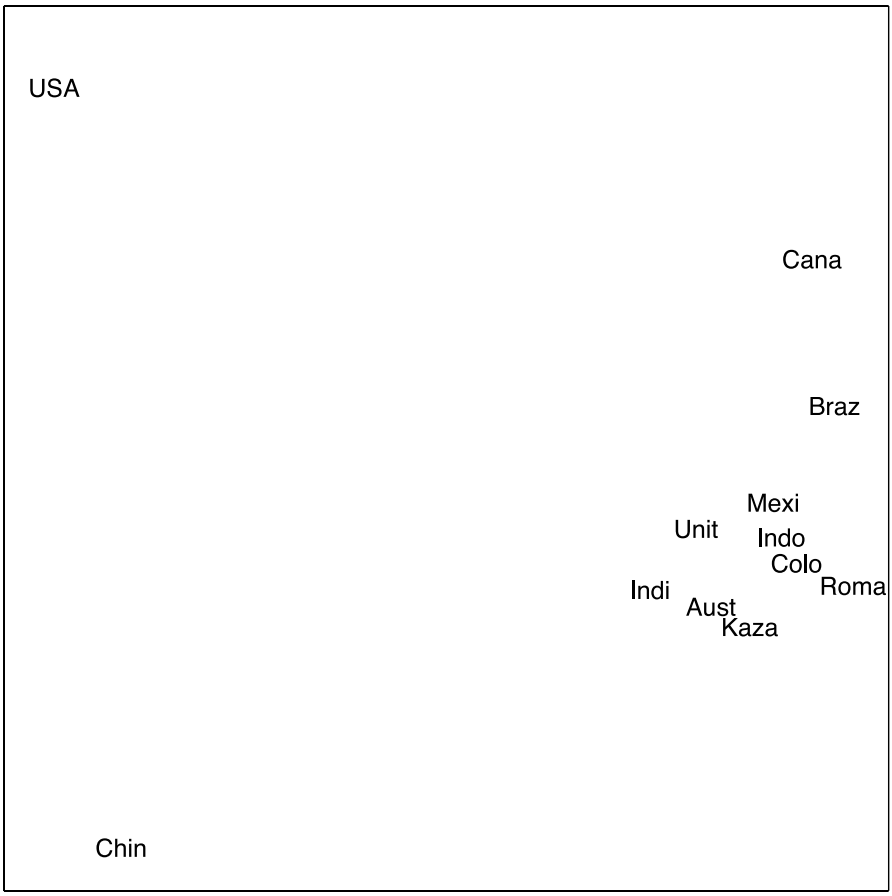

Figure 3.13. INDSCAL for BP data by country: Aust – Australia, Indo – Indonesia, Braz – Brazil, Kaza – Kazakhstan, Cana – Canada, Mexi – Mexico, Chin – China, Roma – Romania, Colo – Colombia, Unit – United Kingdom, Indi – India, USA – USA

Table 3.6. Energy source codes for the BP data

Energy source	code
Coal – Consumption in millions of tonnes of oil equivalent	ColC
Coal – Production in millions of tonnes of oil equivalent	ColP
Hydro – Consumption in millions of tonnes of oil equivalent	Hydr
Natural Gas – Consumption in millions of tonnes of oil equivalent	GasC
Natural Gas – Production in millions of tonnes of oil equivalent	GasP
Nuclear – Consumption in millions of tonnes of oil equivalent	NucC
Oil – Consumption in millions of tonnes	OilC
Oil – Production in millions of tonnes	OilP

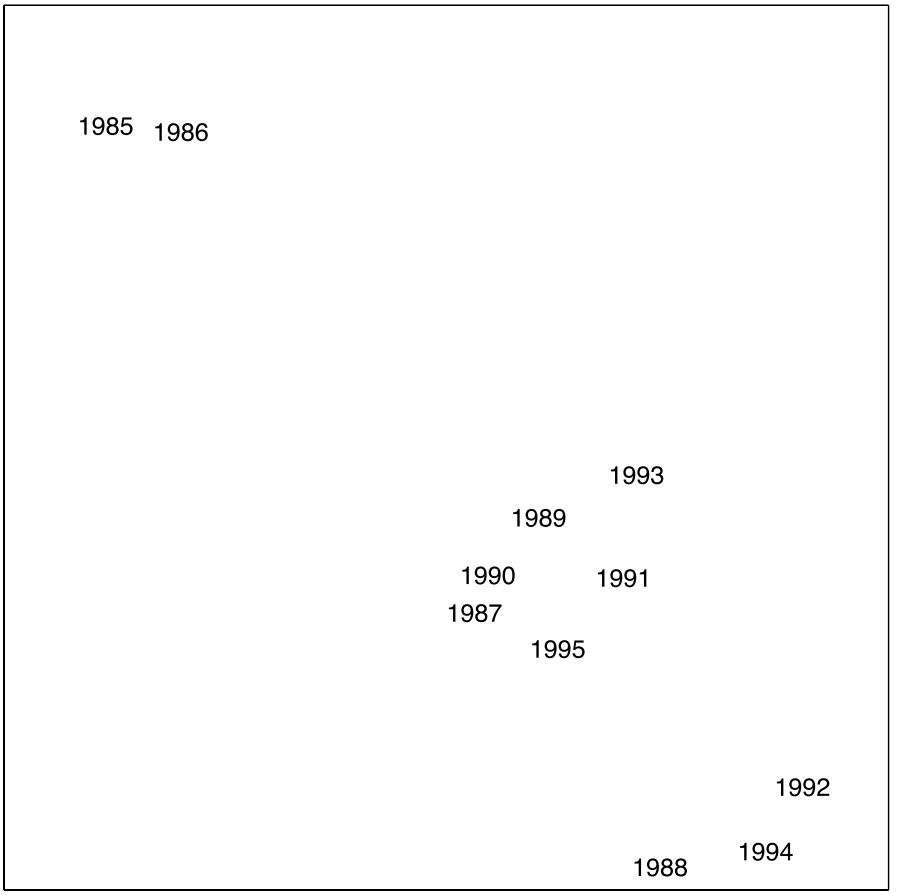

Figure 3.14. INDSCAL for BP data by year

which best represent the corresponding dissimilarities for that individual. Hence the acronym INDSCAL – INdividual Differences SCALing.

Let there be n objects under study and N subjects producing the dissimilarities. Let the dimension of the spaces be p and the points in the group stimulus space be denoted by x_{rt} $(r = 1, \ldots, n;\ t = 1, \ldots, p)$. Let the dissimilarity between objects r and s for the ith subject be $\delta_{rs,i}$ and the points in the subjects' space have coordinates w_{it} $(i = 1, \ldots, N;\ t = 1, \ldots, p)$. Then the weighted Euclidean distance between the rth and sth points for the ith subject is

$$d_{rs,i} = \left[\sum_{t=1}^{p} w_{it} (x_{rt} - x_{st})^2 \right]^{1/2}. \tag{3.10}$$

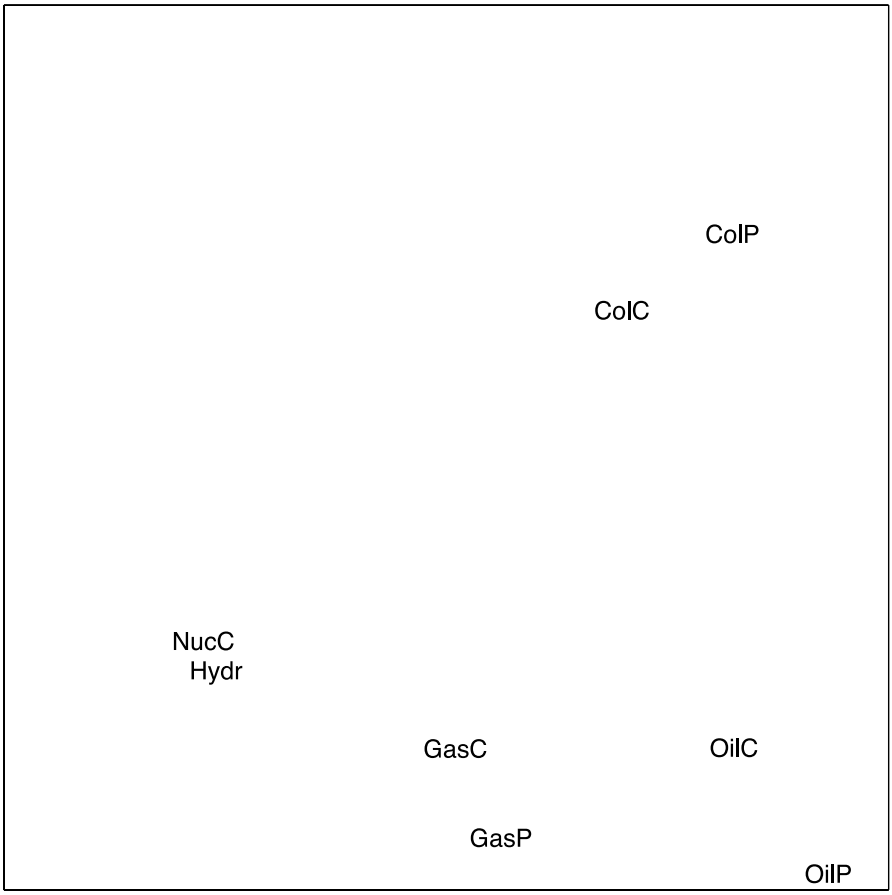

Figure 3.15. INDSCAL for BP data by data source

The individual weights $\{w_{it}\}$ and coordinates $\{x_{rt}\}$ are sought which best match $\{d_{rs,i}\}$ to $\{\delta_{rs,i}\}$. Carroll and Chang (1970) give an algorithm which uses a recursive least-squares approach to do this.

The data used to illustrate INDSCAL are from the 1995 edition of the BP Statistical Review of World Energy. The review incorporates additional elements from the BP Review of World Gas. The review is a compendium of statistics on the primary forms of energy (BP 1996).

The data are for all years from 1985 to 1995, with energy sources as shown in Table 3.6. Data are available for both production and consumption.

Initially dissimilarities were generated for countries and each year, averaging over the energy sources. The INDSCAL analysis results are given in Figs. 3.13 and 3.14. Figure 3.13 shows the "group stimulus space" and Fig. 3.14 the "subjects space".

Clearly China and the USA are exceptional. The USA is the largest consumer/producer of gas and oil, and also the largest consumer of nuclear. In coal (both produc-

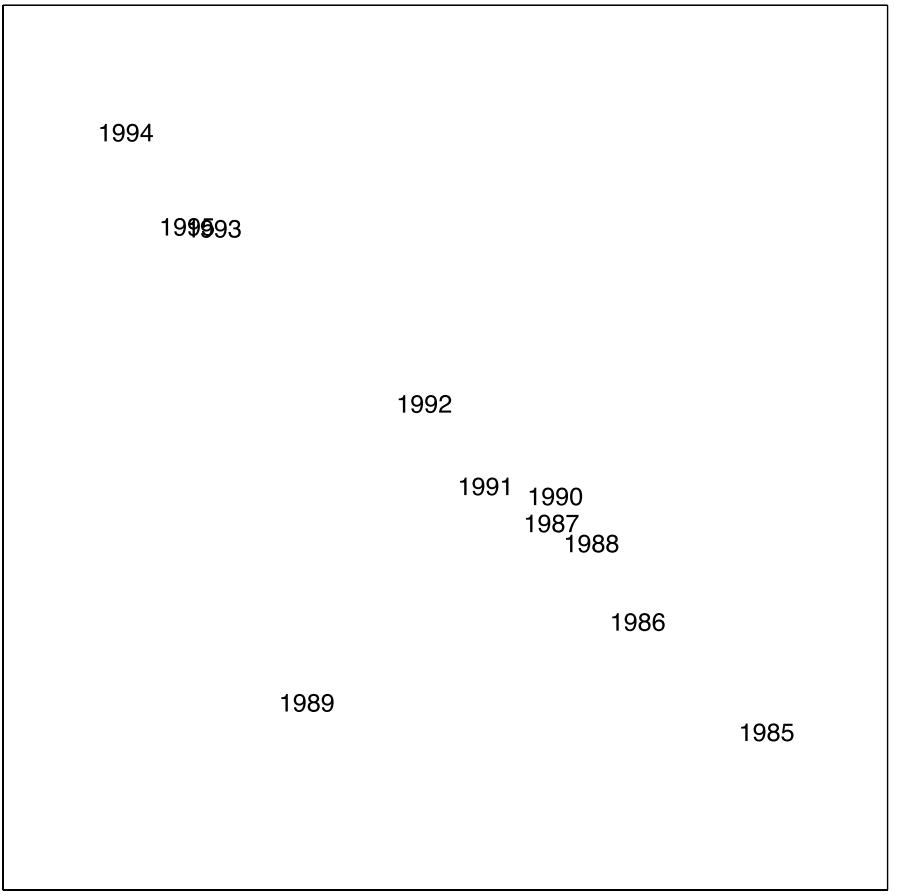

Figure 3.16. Example plots from INDSCAL for BP data by year

tion and consumption) the USA and China are very close and significantly higher than the other countries considered.

The years 1985 and 1986 are exceptional; these correspond to a marked percentage increase in the consumption/production of coal by Indonesia.

A similar analysis may be conducted based on averaging over the countries. The resulting plots are shown in Figs. 3.15 and 3.16.

Clearly the production and consumption of each energy source are very close in the plot showing consumption is highly linked to production. What is surprising is the coincidence of nuclear and hydroelectric energy.

A Procrustes statistic was employed to compare Figs. 3.14 and 3.16, giving a value of 0.59, suggesting the plots are dissimilar. In Fig. 3.16 the years, with slight adjustments, are in sequential order 1994, 1995, 1993, 1992, 1991, 1990, 1987, 1988, 1986, 1985, the exception being 1989.

3.8 Correspondence Analysis and Reciprocal Averaging

Correspondence analysis represents the rows and columns of a two-way contingency table as points in Euclidean spaces of dimension usually chosen as two. It can also be used on any data matrix that has non-negative entries and can also be extended to higher way tables, but this aspect is not covered here. The technique will be illustrated using the contingency table displayed in Table 3.7, which records the number of papers in a research database which employ specific keywords in their descriptive fields (title, keywords or abstract) over an 11-year period.

First distances are measured between the rows of the table. Ordinary Euclidean distance, treating the 11 columns as 11 dimensions for the rows, is not appropriate, since doubling the size of the sample for any one row will alter the Euclidean distance dramatically. For contingency tables where only the overall total is fixed, this may be not be a problem, but it will be for tables where row totals can be chosen arbitrarily. To overcome this problem, χ^2 distances are used. The χ^2 distance, $d_{ii'}$, between the ith and i'th rows is defined by

$$d_{ii'}^2 = \sum_{j=1}^{p} \frac{1}{c_j} \left(\frac{x_{ij}}{r_i} - \frac{x_{i'j}}{r_{i'}} \right)^2 , \tag{3.11}$$

where x_{ij} is the entry in the table in the ith row and jth column, r_i is the ith row sum and c_j is the jth column sum. A space is now found where points in the space represent the rows of the table and where Euclidean distance between points equals the χ^2 distance between the corresponding rows of the table. Greenacre (1984) gives a comprehensive account of how this space is found, but only a brief summary can be given here.

Let $\mathcal{X}$ denote the table that has been normalized to have overall total equal to unity. Let $\mathcal{D}_r$ be the diagonal matrix of the row sums of $\mathcal{X}$, and let $\mathcal{D}_c$ be the diagonal matrix of column sums of $\mathcal{X}$. Let the generalized singular value decomposition of $\mathcal{X}$ be given by

$$\mathcal{X} = \mathcal{A}\mathcal{D}_\lambda \mathcal{B}^T , \tag{3.12}$$

Table 3.7. References employing keywords

Keywords	94	95	96	97	98	99	00	01	02	03	04	Total
Reciprocal av.	0	0	4	0	3	3	0	0	0	3	3	**16**
Correspond. anal.	144	171	186	219	237	246	241	243	278	314	310	**2589**
Ind. diff. scal.	8	4	6	3	5	4	4	4	3	1	3	**45**
Classical scaling	5	6	7	5	2	10	7	7	4	5	6	**64**
Procrustes anal.	15	16	29	20	36	43	41	32	40	70	67	**409**
Mult. scaling	75	117	125	107	109	137	145	147	166	191	195	**1514**
Total	**247**	**314**	**357**	**354**	**392**	**443**	**438**	**433**	**491**	**584**	**584**	**4637**

where $\mathcal{A}^T\mathcal{D}_r^{-1}\mathcal{A} = \mathcal{B}^T\mathcal{D}_c^{-1}\mathcal{B} = \mathcal{I}$. Then the space and coordinates for the row points are given by $\mathcal{D}_r^{-1}\mathcal{A}\mathcal{D}_\lambda$. Note there is a "trivial dimension" which has to be ignored which corresponds to the singular value of unity with singular vector a vector of ones. The Euclidean distances between the points in the $\mathcal{D}_r^{-1}\mathcal{A}\mathcal{D}_\lambda$ space equal the corresponding χ^2 distances between the rows of the table. However, this space has dimension equal to one less than the number of columns. As an approximation, only the first two singular values and corresponding singular vectors are used (ignoring the trivial dimension). Similarly, a space for the columns can be found as $\mathcal{D}_c^{-1}\mathcal{B}\mathcal{D}_\lambda$, and distances between points in this space equal the χ^2 distances between corresponding columns in the table.

Figure 17 shows the space for the rows of the contingency table, and Fig. 3.18 the space for the columns of the table.

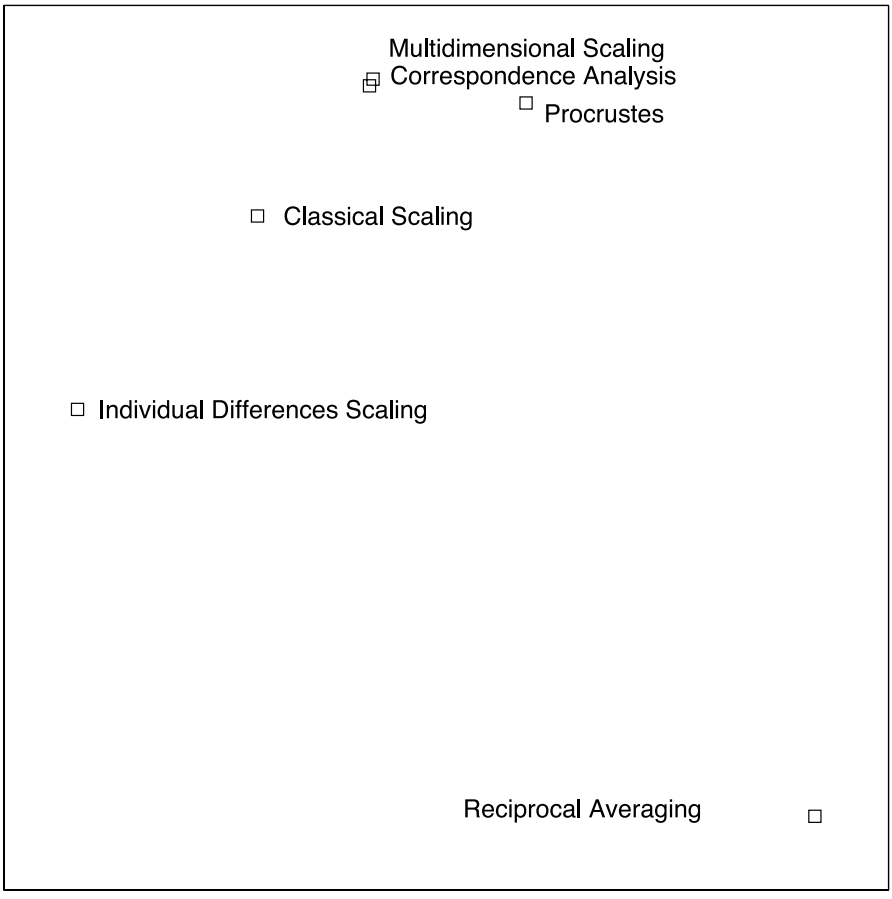

Figure 3.17. Keyword plot for references

The keyword (row) plot shows the similar popularity of correspondence analysis and multidimensional scaling and how Procrustes analysis is related to them, while the difference in use of reciprocal averaging, which is used sporadically, becomes clear. Sometimes the row and column spaces are combined into one space since they have both arisen from the singular value decomposition of $\mathcal{X}$. Distances between row points and column points are not defined, although row points close to column points will have some association.

The year plot is consistent with a steady increase in the use of correspondence analysis, Procrustes analysis and multidimensional scaling. Note that the years to the lower right of the plot are the years that reciprocal averaging makes an entry into the table.

Reciprocal averaging is used for binary data and is essentially the same as correspondence analysis, although the construction appears different. It can be easily explained using an ecological example. Suppose n different species of plants are each

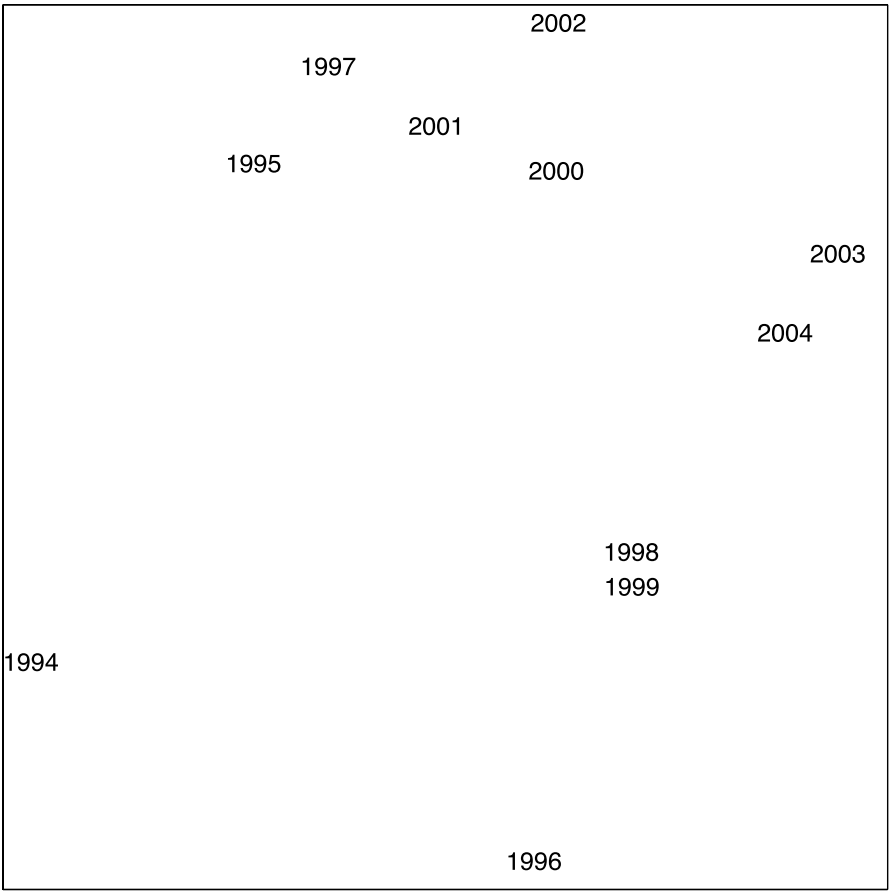

Figure 3.18. Year plot for references

planted at p different sites which vary in exposure to the weather. Let $x_{ij} = 1$ if the ith species survives at the jth site, and zero if it does not. Let u_i be a hardiness score for the ith species and v_j an exposure score for the jth site. It is assumed that the exposure score at the jth site is proportional to the mean hardiness score of the species at that site. Thus

$$v_j \propto \sum_i u_i x_{ij} / \sum_i x_{ij} \,. \tag{3.13}$$

Similarly, it is assumed that the hardiness score of species i is proportional to the mean exposure score of the sites occupied by that species. Thus

$$u_j \propto \sum_j v_j x_{ij} / \sum_j x_{ij} \,. \tag{3.14}$$

Let $r_i = \sum_j x_{ij}$, $c_j = \sum_i x_{ij}$. Reciprocal averaging solves the equations

$$\rho u_i = \sum_j v_j x_{ij} / r_i \quad (i = 1, \ldots, n) \,, \tag{3.15}$$

$$\rho v_j = \sum_i u_i x_{ij} / c_j \quad (j = 1, \ldots, p) \,, \tag{3.16}$$

where ρ is a scaling parameter, to obtain the hardiness and exposure scores.

Reciprocal averaging was used on the Shakespeare data used in Sect. 4, but where it has been turned into a binary format scoring 0/1 if a word is absent/present in each of the plays. The resulting plots are shown in Figs. 3.19 and 3.20.

As in Fig. 3.6 it is the personal words (princess, bishop, clown, moor etc.) that convey most information about the play.

When examining the plays, losing the detail (the word count for each play) has clearly affected the detail displayed, although certain similarities within the plots can be seen, for instance the cluster asy, cym, tam, tem, mer seen in Fig. 3.20 is within a larger cluster within Fig. 3.7.

Large Data Sets and Other Numerical Approaches

3.9

Behind most MDS techniques there is a need for accurate and efficient algorithms for minimizing functions, but many MDS programs and algorithms cannot cope with very large data sets, as they suffer from high computational complexity. They cannot feasibly be applied to data sets over a few thousand objects in size. However, methods have been proposed to overcome this problem, for example Fast Spring Model–Visualisation (FSMvis). FSMvis adopts a novel hybrid approach based upon stochastic sampling, interpolation, and spring models. Following Morrison et al. (2003) the mechanics of the spring model are described.

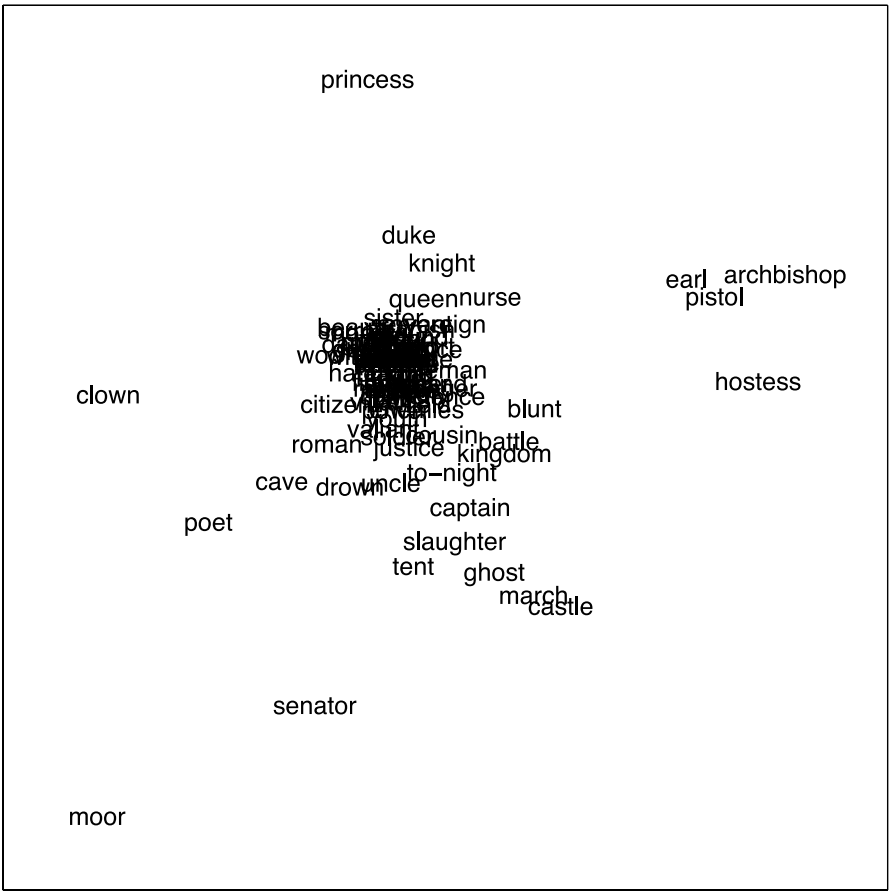

Figure 3.19. Reciprocal averaging of Shakespearean keywords

The concept of the "spring model" comes from work by Eades (1984) on a heuristic approach for graph drawing. Eades described an approach for the layout of a general graph through the physical analogy of a system of steel rings connected by springs. A graph consists of vertices and edges, and here each vertex represents one of the objects under consideration. The graph may be represented by a mechanical system on replacing the vertices by rings and the edges by springs. Each relaxed spring length or "rest distance" is set to be the dissimilarity measured between the corresponding objects. Initially the vertices/rings are placed in random positions and the springs connecting them are either stretched or compressed. When the system is released, the forces exerted by the springs move the system to equilibrium, and presumably to a state of minimal energy. The algorithm employed is iterative, with each iteration refining the layout of the graph.

This procedure was applied to dissimilarities calculated for a population of women who were at least 21 years old, of Pima Indian heritage and living near Phoenix, AZ,

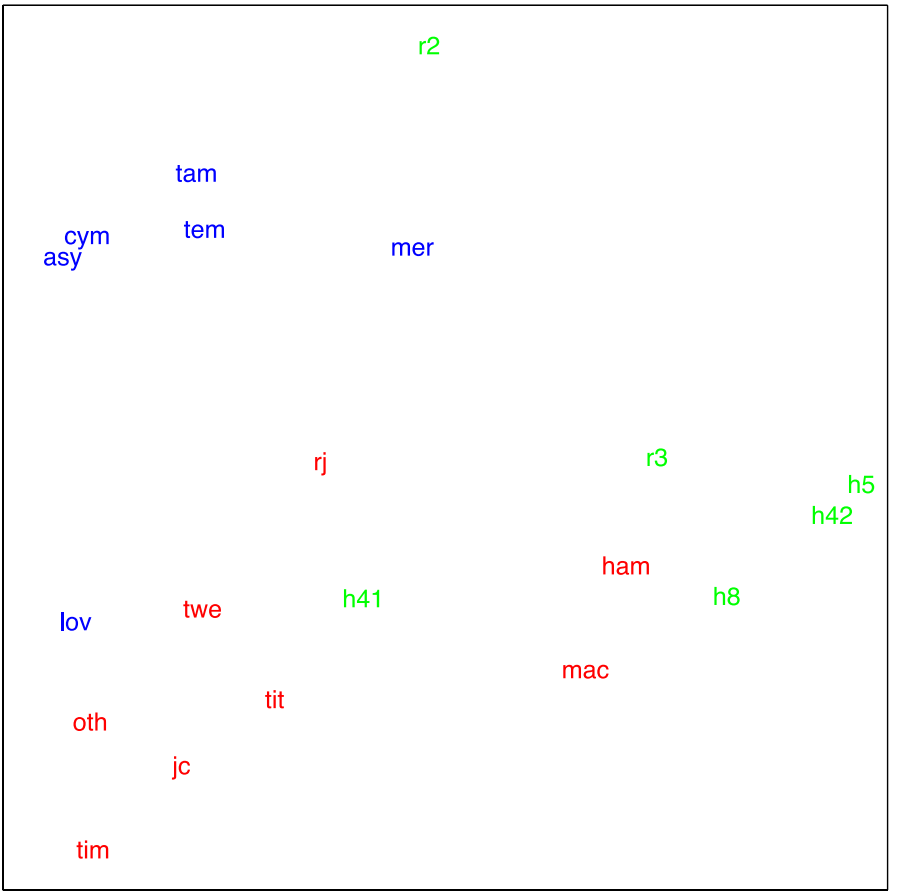

Figure 3.20. Reciprocal averaging of Shakespearean plays

and who were tested for diabetes according to World Health Organization criteria. The results are displayed in Fig. 3.21. The data were collected by the US National Institute of Diabetes and Digestive and Kidney Diseases (Hettich et al., 1998). The raw data consisted of 8 measurements on 768 individuals. The method successfully partitions the 768 individuals into the 2 groups.

Agrafiotis et al. (2001) present a family of algorithms that combine non-linear mapping techniques using neural networks. The method employs an algorithm to project a small random sample and then "learns" the underlying transform using one or more multilayer perceptrons. This approach captures the non-linear mapping relationship as an explicit function and allows the scaling of additional patterns as they become available, without the need to reconstruct the entire map. The approach is particularly useful for extracting low-dimensional Cartesian coordinate vectors from large binary spaces, such as those encountered in the analysis of large chemical data sets. Molecular similarity is used to analyse chemical phenomena and can

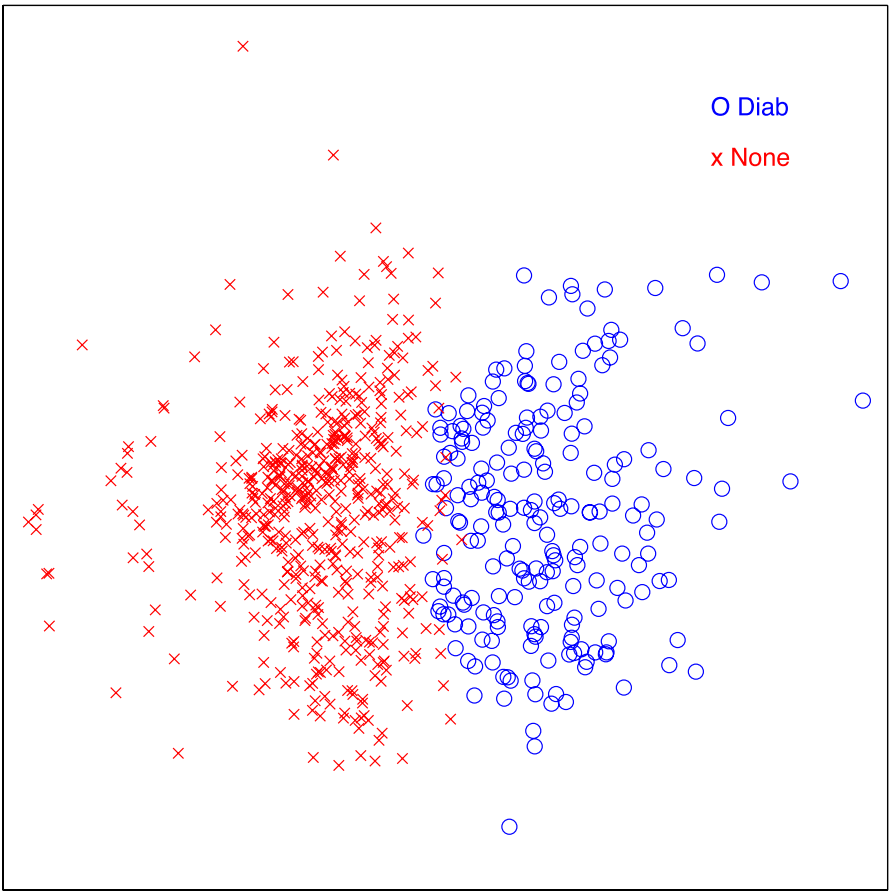

Figure 3.21. Example plot from FSMvis for Pima Indian data

aid in the design of new chemical entities with improved physical, chemical and biological properties. As a guide the authors report analysing one billion items in around 4 h.

Genetic algorithms are an approach to optimization suggested by the biological process of evolution driven by natural selection. The aim is to derive a parameter set that minimizes the difference between a model's expected values and those observed from the data. For detailed reviews see Charbonneau (1995) and Schmitt (2001). The procedure employed for our comparisons uses the algorithm developed by Charbonneau and Knapp (2005). As expected, when tested on the rail journey cost data, the results were indistinguishable. In principle this approach admits larger data sets, but not necessarily of sufficient size to make the approach worthwhile.

Simulated annealing is a suitable approach for large-scale optimization problems. It is claimed to be ideal for locating an ideal global minimum located among a number of local minima. This method has been employed to address the famous travelling

salesman problem. The approach borrows its philosophy from theoretical physics. By analogy the system is raised to a high temperature in which the individuals move freely with respect to each other. As cooling occurs this mobility is lost. Some alignment occurs between the individuals, and the system approaches a minimum energy state. To try to achieve the desired global minimum, the cooling must occur slowly. This procedure has been encapsulated within published algorithms (Goffe et al., 1994; Corana et al., 1987). The approach differs from the conventional gradient descent method in that an occasional step away from an apparent minimum might be used to assist in escaping from local minima.

The majorization approach to minimization was first proposed by de Leeuw (1977) for use with MDS. Essentially, a complicated function $f(x)$ is replaced by a more manageable auxiliary function $g(x, y)$ such that for each x in the domain of f, $f(x) \leq g(x, y)$, for a particular y in the domain of g, and also so that $f(y) = g(y, y)$. The function g is called the majorizing function. An initial value x_0 is used and then $g(x, x_0)$ is minimized with respect to x. Let the value of x, which gives rise to the minimum, be x_1. Then $g(x, x_1)$ is minimized with respect to x, and so on until convergence.

References

Agrafiotis, D.K., Rassokhin, D.N. and Lobanov, V.S. (2001). *J Comp Chem* 22:488–500

Anderberg, M.R. (1973). *Cluster Analysis for Applications*. Academic, New York

Baulieu, F.B. (1989). *J Classificat* 6:233–246

Borg, I. and Groenen, P.G. (1997). *Modern Multidimensional Scaling*. Springer, New York

BP (1996). http://www.bp.com

Carroll, J.D. and Chang, J.J. (1970) Analysis of individual differences in multidimensional scaling via an *n*-way generalization of "Eckart–Young" decomposition. *Psychometrika* 35:283–319

Charbonneau, P. (1995). *Astrophys J Suppl Ser* 101:309–334

Charbonneau, P. and Knapp, B. (2005). http://downloadhaoucaredu/archive/pikaia/

Corana, A., Marchesi, M., Martini, C. and Ridella, S. (1987). *ACM Trans Math Softw* 13:262–280

Cormack, R.M. (1971). *J R Stat Soc A* 134:321–367

Cox, T.F. (2005). *An Introduction to Multivariate Data Analysis*. Hodder Arnold, London

Cox, T.F. and Cox, M.AA. (2001). *Multidimensional Scaling*. Chapman & Hall/CRC, Boca Raton, FL

de Leeuw, J. (1977). Applications of convex analysis to multidimensional scaling. In: Barra, J.R., Brodeau, F., Romier, G., van Cutsen, B. (eds) *Recent Developments in Statistics*. North Holland, Amsterdam

Diday, E. and Simon, J.C. (1976). Clustering analysis. In: Fu, K.S. (ed) *Communication and Cybernetics 10 Digital Pattern Recognition*. Springer, Berlin Heidelberg New York

Digby, P.GN. and Kempton, R.A. (1987). *Multivariate Analysis of Ecological Communities*. Chapman and Hall/CRC, London

Eades, P. (1984). *Congressus Numerantium* 42:149–160

Eslava-Gomez, G. (1989). *Projection pursuit and other graphical methods for multivariate Data*. DPhil Thesis, University of Oxford, Oxford

Everitt, B.S. and Dunn, G. (1983). *Advanced Methods of Data Exploration and Modelling*. Heinemann, London

Fauquet, C., Desbois, D., Fargette, D. and Vidal, G. (1988). Classification of furoviruses based on the amino acid composition of their coat proteins. In: Cooper, J.I., Asher, M.J.C. (eds) *Viruses with Fungal Vectors*. Association of Applied Biologists, Edinburgh

Goffe, W.L., Ferrier, G.D. and Rogers, J. (1994). *J Econometr* 60:65–99

Gordon, A.D. (1999). *Classification*, 2nd edn. Chapman and Hall/CRC, London

Gower, J.C. (1971). *Biometrics* 27:857–874

Gower, J.C. (1985). Measures of similarity, dissimilarity and distance. In: Kotz, S, Johnson, N.L., Read, C.B. (eds) *Encyclopedia of Statistical Sciences*. vol 5. Wiley, New York

Gower, J.C. and Legendre, P. (1986). *J Classificat* 3:5–48

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

Greenacre, M.J. (1984). *Theory and Application of Correspondence Analysis*. Academic, London

Guttman, L. (1968). *Psychometrika* 33:469–506

Hettich, S., Blake, C.L. and Merz, C.J. (1998). *UCI Repository of machine learning databases*. http://www.ics.uci.edu/~mlearn/MLRepository.html

Hubalek, Z. (1982). *Biol Rev* 57:669–689

Hubert, L. and Arabie, P. (1986). Unidimensional scaling and combinatorial optimisation. In: de Leeuw, J., Heiser, W.J., Meulman, J., Critchley, F. (eds) *Multidimensional Data Analysis*. DSWO, Leiden

Hubert, L. and Arabie, P. (1988). Relying on necessary conditions for optimization: unisdimensional scaling and some extensions. In: Bock, H.H. (ed) *Classification and Related Methods of Data Analysis*. North Holland, Amsterdam

Jackson, D.A., Somers, K.M. and Harvey, H.H. (1989). *Am Nat* 133:436–453

Jardine, N. and Sibson, R. (1971). *Mathematical Taxonomy*. Wiley, London

Kruskal, J.B. (1964a). *Psychometrika* 29:1–27

Kruskal, J.B. (1964b). *Psychometrika* 29:115–129

Lau, K.N., Leung, P.L. and Tse, K.K. (1998). *J Classificat* 15:3–14

Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979). *Multivariate Analysis*. Academic, London

Morrison, A., Ross, G. and Chalmers, M. (2003). *Inf Visual* 2:68–77

Ripley, B.D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge

Sammon, J.W. (1969). *IEEE Trans Comput* 18:401–409

Schmitt, L.M. (2001). *Theor Comput Sci* 259:1–61

Shepard, R.N. (1962a). *Psychometrika* 27:125–140

Shepard, R.N. (1962b). *Psychometrika* 27:219–246
Sneath, P. and Sokal, R. (1973). *Numerical Taxonomy*. Freeman, San Francisco
Snijders, T.A.B., Dormaar, M., van Schuur, W.H., Dijkman-Caes, C. and Driessen, G. (1990). *J Classificat* 7:5–31
Torgerson, W.S. (1952). *Psychometrika* 17:401–419
Young, G. and Householder, A.S. (1938). *Psychometrika* 3:19–22
Young, F.W. and Hamer, R.M. (eds) (1987). *Multidimensional Scaling: History, Theory and Applications*. Lawrence Erlbaum, Hillsdale, NJ