14
Multidimensional Scaling

14.1 Introduction

Multidimensional scaling (MDS) is concerned with the problem of constructing a configuration of n points in Euclidean space using information about the distances between the n objects. The interpoint distances themselves may be subject to error. Note that this technique differs from those described in earlier chapters in an important way. Previously, the data points were directly observable as n points in p-space, but here we can only observe a function of the data points, namely the $\frac{1}{2}n(n-1)$ distances.

A simple test example is given in Table 14.1.1, where we are given the road distances (not the "shortest distances") between towns, and the aim is to construct a geographical map of England based on this information. Since these road distances equal the true distances subject to small perturbations, we expect that any sensible MDS method will produce a configuration which is "close" to the true map of these towns.

However, the distances need not be based on Euclidean distances, and can represent many types of dissimilarities between objects. Also in some cases, we start not with dissimilarities but with a set of *similarities* between objects. (We have already given a general account of different types of distances and similarities in Section 13.4.) An example of similarity between two Morse code signals could be the percentage of people who think the Morse code sequences corresponding to the pair of characters are identical after hearing them in rapid succession. Table 14.1.2 gives such data for characters consisting of the 10 numbers in Morse code. These "similarities" can then be used to plot the signals in two-dimensional space. The purpose of this plot is to observe which signals were "like", i.e. near, and which were "unlike", i.e. far from each other, and also to observe the general interrelationship between signals.

Table 14.1.1 Road distances in miles between 12 British towns †

	1	2	3	4	5	6	7	8	9	10	11	12
1												
2	244											
3	218	350										
4	284	77	369									
5	197	167	347	242								
6	312	444	94	463	441							
7	215	221	150	236	279	245						
8	469	583	251	598	598	169	380					
9	166	242	116	257	269	210	55	349				
10	212	53	298	72	170	392	168	531	190			
11	253	325	57	340	359	143	117	264	91	273		
12	270	168	284	164	277	378	143	514	173	111	256	

† 1 = Aberystwyth, 2 = Brighton, 3 = Carlisle, 4 = Dover, 5 = Exeter, 6 = Glasgow, 7 = Hull, 8 = Inverness, 9 = Leeds, 10 = London, 11 = Newcastle, 12 = Norwich.

Definition *An* $(n \times n)$ *matrix* $\mathbf{D}$ *is called a distance matrix if it is symmetric and*

$$d_{rr} = 0, \qquad d_{rs} \geq 0, \qquad r \neq s.$$

Starting with a distance matrix $\mathbf{D}$, the object of MDS is to find points $P_1, \ldots, P_n$ in k dimensions such that if $\hat{d}_{rs}$ denotes the Euclidean distance between P_r and P_s, then $\hat{\mathbf{D}}$ is "similar" in some sense to $\mathbf{D}$. The points P_r are unknown and usually the dimension k is also unknown. In practice

Table 14.1.2 Percentage of times that the pairs of Morse code signals for two numbers were declared to be the same by 598 subjects (Rothkopf, 1957; the reference contains entries for 26 letters as well)

	1	2	3	4	5	6	7	8	9	0
1	84									
2	62	89								
3	16	59	86							
4	6	23	38	89						
5	12	8	27	56	90					
6	12	14	33	34	30	86				
7	20	25	17	24	18	65	85			
8	37	25	16	13	10	22	65	88		
9	57	28	9	7	5	8	31	58	91	
0	52	18	9	7	5	18	15	39	79	94

one usually limits the dimension to $k = 1, 2$, or 3 in order to facilitate the interpretation of the solution.

Nature of the solution It is important to realize that the configuration produced by any MDS method is indeterminate with respect to translation, rotation, and reflection. In the two-dimensional case of road distances (Table 14.1.1), the whole configuration of points can be "shifted" from one place in the plane to another and the whole configuration can be "rotated" or "reflected".

In general, if $P_1, \ldots, P_n$ with coordinates $\mathbf{x}_i' = (x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$, represents an MDS solution in p dimensions, then

$$\mathbf{y}_i = \mathbf{A}\mathbf{x}_i + \mathbf{b}, \qquad i = 1, \ldots, n,$$

is also a solution, where $\mathbf{A}$ is an orthogonal matrix and $\mathbf{b}$ is any vector.

Types of solution Methods of solution using only the rank order of the distances

$$d_{r_1 s_1} < d_{r_2 s_2} < \cdots < d_{r_m s_m}, \qquad m = \tfrac{1}{2}n(n-1), \tag{14.1.1}$$

where $(r_1, s_1), \ldots, (r_m, s_m)$ denotes all pairs of subscripts of r and s, $r < s$, are termed *non-metric methods of multidimensional scaling.*

The rank orders are invariant under monotone increasing transformations f of the d_{rs}, i.e.

$$d_{r_1 s_1} < d_{r_2 s_2} < \cdots \Leftrightarrow f(d_{r_1 s_1}) < f(d_{r_2 s_2}) < \cdots.$$

Therefore the configurations which arise from non-metric scaling are indeterminate not only with respect to translation, rotation, and reflection, but also with respect to uniform expansion or contraction.

Solutions which try to obtain P_i directly from the given distances are called *metric methods.* These methods derive P_r such that, in some sense, the new distances $\hat{d}_{rs}$ between points P_r and P_s are as close to the original d_{rs} as possible.

In general the purpose of MDS is to provide a "picture" which can be used to give a meaningful interpretation of the data. Hopefully, the picture will convey useful information about the relationships between the objects. Note that this chapter differs from most of the earlier chapters in that no probabilistic framework is set up; the technique is purely data-analytic.

One important use of MDS is seriation. The aim here is to order a set of objects chronologically on the basis of dissimilarities or similarities between them. Suppose the points in the MDS configuration in $k > 2$ dimensions lie nearly on a smooth curve. This property then suggests that the differences in the data are in fact one-dimensional and the ordering of

the points along this curve can be used to seriate the data. (See D. G. Kendall, 1971.)

14.2 Classical Solution

14.2.1 Some theoretical results

Definition *A distance matrix* $\mathbf{D}$ *is called* Euclidean *if there exists a configuration of points in some Euclidean space whose interpoint distances are given by* $\mathbf{D}$*; that is, if for some* p*, there exists points* $\mathbf{x}_1, \ldots, \mathbf{x}_n \in R^p$ *such that*

$$d_{rs}^2 = (\mathbf{x}_r - \mathbf{x}_s)'(\mathbf{x}_r - \mathbf{x}_s). \tag{14.2.1}$$

The following theorem enables us to tell whether $\mathbf{D}$ is Euclidean, and, if so, how to find a corresponding configuration of points. First we need some notation. For any distance matrix $\mathbf{D}$, let

$$\mathbf{A} = (a_{rs}), \qquad a_{rs} = -\tfrac{1}{2}d_{rs}^2 \tag{14.2.2}$$

and set

$$\mathbf{B} = \mathbf{HAH}, \tag{14.2.3}$$

where $\mathbf{H} = \mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}'$ is the $(n \times n)$ centring matrix.

Theorem 14.2.1 *Let* $\mathbf{D}$ *be a distance matrix and define* $\mathbf{B}$ *by* (14.2.3)*. Then* $\mathbf{D}$ *is Euclidean if and only if* $\mathbf{B}$ *is p.s.d. In particular, the following results hold:*

(a) *If* $\mathbf{D}$ *is the matrix of Euclidean interpoint distances for a configuration* $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)'$*, then*

$$b_{rs} = (\mathbf{z}_r - \bar{\mathbf{z}})'(\mathbf{z}_s - \bar{\mathbf{z}}), \qquad r, s = 1, \ldots, n. \tag{14.2.4}$$

In matrix form (14.2.4) *becomes* $\mathbf{B} = (\mathbf{HZ})(\mathbf{HZ})'$ *so* $\mathbf{B} \geq 0$*. Note that* $\mathbf{B}$ *can be interpreted as the "centred inner product matrix" for the configuration* $\mathbf{Z}$*.*

(b) *Conversely, if* $\mathbf{B}$ *is p.s.d. of rank* p *then a configuration corresponding to* $\mathbf{B}$ *can be constructed as follows. Let* $\lambda_1 > \cdots > \lambda_p$ *denote the positive eigenvalues of* $\mathbf{B}$ *with corresponding eigenvectors* $\mathbf{X} = (\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(p)})$ *normalized by*

$$\mathbf{x}_{(i)}'\mathbf{x}_{(i)} = \lambda_i, \qquad i = 1, \ldots, p. \tag{14.2.5}$$

Then the points P_r *in* R^p *with coordinates* $\mathbf{x}_r = (x_{r1}, \ldots, x_{rp})'$ *(so* $\mathbf{x}_r$ *is the* r*th* row *of* $\mathbf{X}$*) have interpoint distances given by* $\mathbf{D}$*. Further, this*

configuration has centre of gravity $\bar{\mathbf{x}} = \mathbf{0}$, *and* $\mathbf{B}$ *represents the inner product matrix for this configuration.*

Proof We first prove (a). Suppose

$$d_{rs}^2 = -2a_{rs} = (\mathbf{z}_r - \mathbf{z}_s)'(\mathbf{z}_r - \mathbf{z}_s). \tag{14.2.6}$$

We can write

$$\mathbf{B} = \mathbf{HAH} = \mathbf{A} - n^{-1}\mathbf{AJ} - n^{-1}\mathbf{JA} + n^{-2}\mathbf{JAJ}, \tag{14.2.7}$$

where $\mathbf{J} = \mathbf{1}\mathbf{1}'$. Now

$$\frac{1}{n}\mathbf{AJ} = \begin{bmatrix} \bar{a}_{1\cdot} & \dots & \bar{a}_{1\cdot} \\ \vdots & & \vdots \\ \bar{a}_{n\cdot} & \dots & \bar{a}_{n\cdot} \end{bmatrix}, \quad \frac{1}{n}\mathbf{JA} = \begin{bmatrix} \bar{a}_{\cdot 1} & \dots & \bar{a}_{\cdot n} \\ \vdots & & \vdots \\ \bar{a}_{\cdot 1} & \dots & \bar{a}_{\cdot n} \end{bmatrix}, \quad \frac{1}{n^2}\mathbf{JAJ} = \begin{bmatrix} \bar{a}_{\cdot\cdot} & \dots & \bar{a}_{\cdot\cdot} \\ \vdots & & \vdots \\ \bar{a}_{\cdot\cdot} & \dots & \bar{a}_{\cdot\cdot} \end{bmatrix},$$

where

$$\bar{a}_{r\cdot} = \frac{1}{n}\sum_{s=1}^{n} a_{rs}, \qquad \bar{a}_{\cdot s} = \frac{1}{n}\sum_{r=1}^{n} a_{rs}, \qquad \bar{a}_{\cdot\cdot} = \frac{1}{n^2}\sum_{r,s=1}^{n} a_{rs}. \tag{14.2.8}$$

Thus

$$b_{rs} = a_{rs} - \bar{a}_{r\cdot} - \bar{a}_{\cdot s} + \bar{a}_{\cdot\cdot}. \tag{14.2.9}$$

After substituting for a_{rs} from (14.2.6) and using (14.2.8), this formula simplifies to

$$b_{rs} = (\mathbf{z}_r - \bar{\mathbf{z}})'(\mathbf{z}_s - \bar{\mathbf{z}}). \tag{14.2.10}$$

(See Exercise 14.2.1 for further details.) Thus (a) is proved.

Conversely, to prove (b) suppose $\mathbf{B} \geqslant 0$ and consider the configuration given in the thorem. Let $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \dots, \lambda_p)$ and let $\mathbf{\Gamma} = \mathbf{X}\mathbf{\Lambda}^{-1/2}$, so that the columns of $\mathbf{\Gamma}$, $\boldsymbol{\gamma}_{(i)} = \lambda_i^{-1/2}\mathbf{x}_{(i)}$ are *standardized* eigenvectors of $\mathbf{B}$. Then by the spectral decomposition theorem (Remark 4 after Theorem A.6.4),

$$\mathbf{B} = \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}' = \mathbf{X}\mathbf{X}';$$

that is, $b_{rs} = \mathbf{x}_r'\mathbf{x}_s$, so $\mathbf{B}$ represents the inner product matrix for this configuration.

We must now show that $\mathbf{D}$ represents the matrix of interpoint distances for this configuration. Using (14.2.9) to write $\mathbf{B}$ in terms of $\mathbf{A}$, we get

$$\begin{aligned} (\mathbf{x}_r - \mathbf{x}_s)'(\mathbf{x}_r - \mathbf{x}_s) &= \mathbf{x}_r'\mathbf{x}_r - 2\mathbf{x}_r'\mathbf{x}_s + \mathbf{x}_s'\mathbf{x}_s \\ &= b_{rr} - 2b_{rs} + b_{ss} \\ &= a_{rr} - 2a_{rs} + a_{ss} \\ &= -2a_{rs} = d_{rs}^2 \end{aligned} \tag{14.2.11}$$

because $a_{rr} = -\frac{1}{2}d_{rr}^2 = 0$ and $-2a_{rs} = d_{rs}^2$.

Finally, note that $\mathbf{B1} = \mathbf{HAH1} = \mathbf{0}$, so that $\mathbf{1}$ is an eigenvector of $\mathbf{B}$ corresponding to the eigenvalue 0. Thus $\mathbf{1}$ is orthogonal to the columns of $\mathbf{X}$, $\mathbf{x}_{(i)}'\mathbf{1} = 0$, $i = 1, \dots, p$. Hence

$$n\bar{\mathbf{x}} = \sum_{r=1}^{n} \mathbf{x}_i = \mathbf{X}'\mathbf{1} = (\mathbf{x}_{(1)}'\mathbf{1}, \dots, \mathbf{x}_{(p)}'\mathbf{1})' = \mathbf{0}$$

so that the centre of gravity of this configuration lies at the origin. ■

Remarks (1) The matrix $\mathbf{X}$ can be visualized in the following way in terms of the eigenvectors of $\mathbf{B}$ and the corresponding points:

		Eigenvalues λ_1 λ_2 … λ_p	Vector notation
Points	P_1	x_{11} x_{12} … x_{1p}	$\mathbf{x}_1'$
	P_2	x_{21} x_{22} … x_{2p}	$\mathbf{x}_2'$
	⋮	⋮ ⋮ ⋮	⋮
	P_n	x_{n1} x_{n2} … x_{np}	$\mathbf{x}_n'$
Vector notation.		$\mathbf{x}_{(1)}\mathbf{x}_{(2)} \dots \mathbf{x}_{(p)}$	

Centre of gravity:

$$\bar{x}_1 = 0, \bar{x}_2 = 0, \dots, \bar{x}_p = 0, \qquad \bar{\mathbf{x}} = \frac{1}{n}\sum \mathbf{x}_r = \mathbf{0}.$$

In short, the rth *row* of $\mathbf{X}$ contains the coordinates of the rth point, whereas the ith *column* of $\mathbf{X}$ contains the eigenvector corresponding to λ_i.

(2) Geometrically, if $\mathbf{B}$ is the centred inner product matrix for a configuration $\mathbf{Z}$, then $b_{rr}^{1/2}$ equals the distance between $\mathbf{z}_r$ and $\bar{\mathbf{z}}$, and $b_{rs}/(b_{rr}b_{ss})^{1/2}$ equals the cosine of the angle subtended at $\bar{\mathbf{z}}$ between $\mathbf{z}_r$ and $\mathbf{z}_s$.

(3) Note that $\mathbf{1}$ is an eigenvector of $\mathbf{B}$ whether $\mathbf{D}$ is Euclidean or not.

(4) The theorem does not hold if $\mathbf{B}$ has negative eigenvalues. The reason can be found in (14.2.5) because it is impossible to normalize a vector to have a negative squared norm.

(5) *History* This result was first proved by Schoenberg (1935) and Young and Householder (1938). Its use as a basis for multidimensional scaling was put forward by Torgerson (1958) and the ideas were substantially amplified by Gower (1966).

14.2.2 A practical algorithm

Suppose we are given a distance matrix $\mathbf{D}$ which we hope can approximately represent the interpoint distances of a configuration in a Euclidean space of low dimension k (usually $k = 1$, 2, or 3). The matrix $\mathbf{D}$ may or may not be Euclidean; however, even if $\mathbf{D}$ is Euclidean, the dimension of the space in which it can be represented will usually be too large to be of practical interest.

One possible choice of configuration in k dimensions is suggested by Theorem 14.2.1. *Choose the configuration in R^k whose coordinates are determined by the first k eigenvectors of* $\mathbf{B}$. If the first k eigenvalues of $\mathbf{B}$ are "large" and positive and the other eigenvalues are near 0 (positive or negative), then hopefully, the interpoint distances of this configuration will closely approximate $\mathbf{D}$.

This configuration is called the *classical solution to the MDS problem in k dimensions.* It is a metric solution and its optimal properties are discussed in Section 14.4. For computational purposes we shall summarize the calculations involved:

(a) From $\mathbf{D}$ construct the matrix $\mathbf{A} = (-\frac{1}{2}d_{rs}^2)$.
(b) Obtain the matrix $\mathbf{B}$ with elements $b_{rs} = a_{rs} - \bar{a}_{r\cdot} - \bar{a}_{\cdot s} + \bar{a}_{\cdot\cdot}$.
(c) Find the k largest eigenvalues $\lambda_1 > \cdots > \lambda_k$ of $\mathbf{B}$ (k chosen ahead of time), with corresponding eigenvectors $\mathbf{X} = (\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(k)})$ which are normalized by $\mathbf{x}_{(i)}'\mathbf{x}_{(i)} = \lambda_i$, $i = 1, \ldots, k$. (We are supposing here that the first k eigenvalues are all positive.)
(d) The required coordinates of the points P_r are $\mathbf{x}_r = (x_{r1}, \ldots, x_{rk})$, $r = 1, \ldots, k$, the *rows* of $\mathbf{X}$.

Example 14.2.1 To illustrate the algorithm, consider a (7×7) distance

matrix

$$\mathbf{D} = \begin{bmatrix} 0 & 1 & \sqrt{3} & 2 & \sqrt{3} & 1 & 1 \\ & 0 & 1 & \sqrt{3} & 2 & \sqrt{3} & 1 \\ & & 0 & 1 & \sqrt{3} & 2 & 1 \\ & & & 0 & 1 & \sqrt{3} & 1 \\ & & & & 0 & 1 & 1 \\ & & & & & 0 & 1 \\ & & & & & & 0 \end{bmatrix}.$$

Constructing the matrix $\mathbf{A}$ from (14.2.2), it is found that

$$\bar{a}_{r\cdot} = -\tfrac{13}{14}, \quad r = 1, \ldots, 6, \quad \bar{a}_{7\cdot} = -\tfrac{3}{7};$$
$$\bar{a}_{r\cdot} = \bar{a}_{\cdot r}, \quad \bar{a}_{\cdot\cdot} = -\tfrac{6}{7}.$$

Hence from (14.2.9) the matrix $\mathbf{B}$ is given by

$$\mathbf{B} = \frac{1}{2}\begin{bmatrix} 2 & 1 & -1 & -2 & -1 & 1 & 0 \\ & 2 & 1 & -1 & -2 & -1 & 0 \\ & & 2 & 1 & -1 & -2 & 0 \\ & & & 2 & 1 & -1 & 0 \\ & & & & 2 & 1 & 0 \\ & & & & & 2 & 0 \\ & & & & & & 0 \end{bmatrix}.$$

The columns of $\mathbf{B}$ are linearly dependent. It can be seen that

$$\mathbf{b}_{(3)} = \mathbf{b}_{(2)} - \mathbf{b}_{(1)}, \quad \mathbf{b}_{(4)} = -\mathbf{b}_{(1)}, \quad \mathbf{b}_{(5)} = -\mathbf{b}_{(2)},$$
$$\mathbf{b}_{(6)} = \mathbf{b}_{(1)} - \mathbf{b}_{(2)}, \quad \mathbf{b}_{(7)} = \mathbf{0}. \tag{14.2.12}$$

Hence the rank of matrix $\mathbf{B}$ is at the most 2. From the leading (2×2) matrix it is clear that the rank is 2. Thus, a configuration exactly fitting the distance matrix can be constructed in $k = 2$ dimensions.

The eigenvalues of $\mathbf{B}$ are found to be

$$\lambda_1 = 3, \quad \lambda_2 = 3, \quad \lambda_3 = \cdots = \lambda_7 = 0.$$

The configuration can be constructed using any two orthogonal vectors

for the eigenspace corresponding to $\lambda = 3$, such as

$$\mathbf{x}'_{(1)} = (a, a, 0, -a, -a, 0, 0), \qquad a = \tfrac{1}{2}\sqrt{3},$$
$$\mathbf{x}'_{(2)} = (b, -b, -2b, -b, b, 2b, 0), \qquad b = \tfrac{1}{2}.$$

Then the coordinates of the seven points are

A	B	C	D	E	F	G
$(\frac{1}{2}\sqrt{3}, \frac{1}{2})$	$(\frac{1}{2}\sqrt{3}, -\frac{1}{2})$	$(0, -1)$	$(-\frac{1}{2}\sqrt{3}, -\frac{1}{2})$	$(-\frac{1}{2}\sqrt{3}, \frac{1}{2})$	$(0, 1)$	$(0, 0)$

The centre of gravity of these points is of course $(0, 0)$, and it can be verified that the distance matrix for these points is **D**. In fact, A to F are vertices of a hexagon with each side of length 1, and the line FC is the x_2 axis. Its centre is G. (Indeed, **D** was constructed with the help of these points.) A similar configuration based on a non-Euclidean distance is described in Exercise 14.2.7.

14.2.3 Similarities

In some situations we start not with distances between n objects, but with similarities. Recall that an $(n \times n)$ matrix **C** is called a similarity matrix if $c_{rs} = c_{sr}$ and if

$$c_{rs} \leq c_{rr} \quad \text{for all } r, s. \tag{14.2.13}$$

Examples of possible similarity matrices were given in Section 13.4.

To use the techniques of the preceding sections, it is necessary to transform the similarities to distances. A useful transformation is the following.

Definition *The* standard transformation *from a similarity matrix* **C** *to a distance matrix* **D** *is defined by*

$$d_{rs} = (c_{rr} - 2c_{rs} + c_{ss})^{1/2}. \tag{14.2.14}$$

Note that if (14.2.13) holds, then the quantity under the square root in (14.2.14) must be non-negative, and that $d_{rr} = 0$. Hence **D** is a distance matrix.

Many of the similarity matrices discussed in Section 13.5 were positive. This property is attractive because the resulting distance matrix, using the standard transformation, is Euclidean.

Theorem 14.2.2 *If* $\mathbf{C} \geqslant 0$, *then the distance matrix* **D** *defined by the standard transformation* (14.2.14) *is Euclidean, with centred inner product matrix* $\mathbf{B} = \mathbf{HCH}$.

Proof First note that since $\mathbf{C} \geqslant 0$,

$$d_{rs}^2 = c_{rr} - 2c_{rs} + c_{ss} = \mathbf{x}'\mathbf{C}\mathbf{x} \geqslant 0,$$

where $\mathbf{x}$ is a vector with $+1$ in the rth place and -1 in the sth place, for $r \neq s$. Thus, the standard transformation is well defined and **D** is a distance matrix.

Let **A** and **B** be defined by (14.2.2) and (14.2.3). Since **HCH** is also p.s.d., it is sufficient to prove that $\mathbf{B} = \mathbf{HCH}$ in order to conclude that **D** is Euclidean with centred inner product matrix **HCH**.

Now $\mathbf{B} = \mathbf{HAH}$ can be written elementwise using (14.2.9). Substituting for $a_{rs} = -\frac{1}{2}d_{rs}^2$ using (14.2.14) gives

$$\begin{aligned} -2b_{rs} &= d_{rs}^2 - \frac{1}{n}\sum_{i=1}^{n} d_{ri}^2 - \frac{1}{n}\sum_{j=1}^{n} d_{js}^2 + \frac{1}{n^2}\sum_{i,j=1}^{n} d_{ij}^2 \\ &= c_{rr} - 2c_{rs} + c_{ss} - \frac{1}{n}\sum_{i=1}^{n}(c_{rr} - 2c_{ri} + c_{ii}) \\ &\quad - \frac{1}{n}\sum_{j=1}^{n}(c_{ij} - 2c_{js} + c_{ss}) + \frac{1}{n^2}\sum_{i,j=1}^{n}(c_{ii} - 2c_{ij} + c_{jj}) \\ &= -2c_{rs} + 2\bar{c}_{r\cdot} + 2\bar{c}_{\cdot s} - 2c_{\cdot\cdot}. \end{aligned}$$

Hence

$$b_{rs} = c_{rs} - \bar{c}_{r\cdot} - \bar{c}_{\cdot s} + c_{\cdot\cdot}$$

or, in matrix form, $\mathbf{B} = \mathbf{HCH}$. Thus the theorem is proved. ■

Example 14.2.2 We now consider the Morse code data given in Table 14.1.2 and described in Section 14.1. The data is presented as a similarity matrix $\mathbf{C} = (c_{rs})$. Using the standard transformation from similarities to distances, take

$$d_{rs} = (c_{rr} + c_{ss} - 2c_{rs})^{1/2}.$$

We obtain the eigenvectors and eigenvalues of **HCH** in accordance with Theorem 14.2.2. It is found that

$$\lambda_1 = 187.4, \quad \lambda_2 = 121.0, \quad \lambda_3 = 95.4, \quad \lambda_4 = 55.4, \quad \lambda_5 = 46.6,$$
$$\lambda_6 = 31.5, \quad \lambda_7 = 9.6, \quad \lambda_8 = 4.5, \quad \lambda_9 = 0.0, \quad \lambda_{10} = -4.1.$$

The first two eigenvectors appropriately normalized are

$$(-4.2, -0.3, 3.7, 5.6, 5.4, 3.8, 0.9, -3.0, -6.2, -5.7),$$
$$(-3.2, -5.8, -4.3, -0.6, 0.0, 4.0, 5.5, 3.6, 0.6, 0.2).$$

However, the first two principal coordinates account for only $100(\lambda_1 + \lambda_2)/\Sigma|\lambda_i|$ percent $= 56\%$ of the total configuration. The points P_r are plotted in Figure 14.2.1. It can be seen that the x_1 axis measures the increasing number of dots whereas the x_2 axis measures the heterogeneity

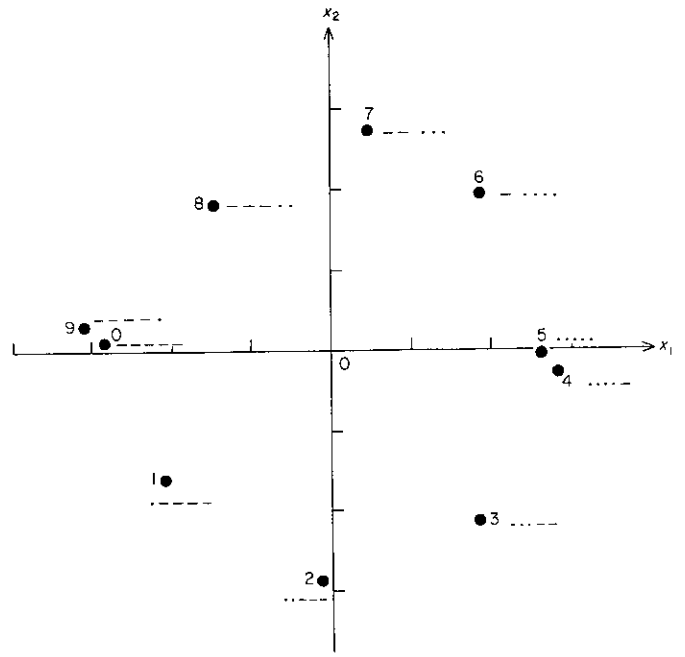


Figure 14.2.1 Classical solution for Morse code data in Table 14.1.1.

of the signal. If we regard the OP_rs as vectors, then angles between consecutive vectors are about 45° except between the vectors from 0 and 9, and 4 and 5. The small separation between these latter points might be expected because the change in just the last character may not make much impact on the untrained ear. Thus the configuration brings out the main features of the general interrelationship between these signals. The points are roughly on a circle in the order $0, 1, \ldots, 9$.

14.3 Duality Between Principal Coordinate Analysis and Principal Component Analysis

So far in this chapter we have treated the $(n \times n)$ matrix $\mathbf{D}$ of distances between n objects as the starting point for our analysis. However, in many situations, we start with a data matrix $\mathbf{X}(n \times p)$ and must make a choice of distance function.

Several possibilities for a distance function were discussed in Section 13.4. The simplest choice is, of course, Euclidean distance. In this case there is a close connection between the work in Section 14.2 and principal component analysis.

Let $\mathbf{X}(n \times p)$ be a data matrix and let $\lambda_1 \geqslant \cdots \geqslant \lambda_p$ be the eigenvalues of $n\mathbf{S} = \mathbf{X}'\mathbf{H}\mathbf{X}$, where $\mathbf{S}$ is the sample covariance matrix. For simplicity we shall suppose the eigenvalues are all non-zero and distinct. Then $\lambda_1, \ldots, \lambda_p$ are also the non-zero eigenvalues of $\mathbf{B} = \mathbf{H}\mathbf{X}\mathbf{X}'\mathbf{H}$. Note that the rows of $\mathbf{H}\mathbf{X}$ are just the centred rows of $\mathbf{X}$, so that $\mathbf{B}$ represents the centred inner product matrix,

$$b_{rs} = (\mathbf{x}_r - \bar{\mathbf{x}})'(\mathbf{x}_s - \bar{\mathbf{x}}).$$

Definition *Let $\mathbf{v}_{(i)}$ be the ith eigenvector of $\mathbf{B}$,*

$$\mathbf{B}\mathbf{v}_{(i)} = \lambda_i \mathbf{v}_{(i)},$$

normalized by $\mathbf{v}_{(i)}'\mathbf{v}_{(i)} = \lambda_i$, $i = 1, \ldots, p$. For fixed k $(1 \leqslant k \leqslant p)$, the rows of $\mathbf{V}_k = (\mathbf{v}_{(1)}, \ldots, \mathbf{v}_{(k)})$ are called the principal coordinates of $\mathbf{X}$ in k dimensions.

Thus, from Theorem 14.2.1, if $\mathbf{D}$ is the Euclidean distance matrix between the rows of $\mathbf{X}$, then the k-dimensional classical solution to the MDS problem is given by the principal coordinates of $\mathbf{X}$ in k dimensions. Principal coordinates are closely linked to principal components, as the following result shows.

Theorem 14.3.1 *The principal coordinates of $\mathbf{X}$ in k dimensions are given by the centred* scores *of the n objects on the first k principal components.*

Proof Let $\boldsymbol{\gamma}_{(i)}$ denote the ith principal component loading vector, standardized by $\boldsymbol{\gamma}_{(i)}'\boldsymbol{\gamma}_{(i)} = 1$, so that by the spectral decomposition theorem (Theorem A.6.4),

$$\mathbf{X}'\mathbf{H}\mathbf{X} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}',$$

where $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_{(1)}, \ldots, \boldsymbol{\gamma}_{(p)})$ and $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_p)$. By the singular value decomposition theorem (Theorem A.6.5), we can choose the signs of $\boldsymbol{\gamma}_{(i)}$ and $\mathbf{v}_{(i)}$, so that $\mathbf{H}\mathbf{X}$ can be written in terms of these eigenvectors as

$$\mathbf{H}\mathbf{X} = \mathbf{V}\boldsymbol{\Gamma}',$$

where $\mathbf{V} = \mathbf{V}_p = (\mathbf{v}_{(1)}, \ldots, \mathbf{v}_{(p)})$.

The scores of the n rows of $\mathbf{H}\mathbf{X}$ on the ith principal component are given by the n elements of $\mathbf{H}\mathbf{X}\boldsymbol{\gamma}_{(i)}$. Thus, writing $\boldsymbol{\Gamma}_k = (\boldsymbol{\gamma}_{(1)}, \ldots, \boldsymbol{\gamma}_{(k)})$, the scores on the first k principal components are given by

$$\mathbf{H}\mathbf{X}\boldsymbol{\Gamma}_k = \mathbf{V}\boldsymbol{\Gamma}'\boldsymbol{\Gamma}_k = \mathbf{V}(\mathbf{I}_k, \mathbf{0})' = \mathbf{V}_k,$$

since the columns of Γ are orthogonal to one another. Hence the theorem is proved. ■

Since the columns of Γ_k are orthogonal to one another, $\Gamma_k'\Gamma_k = \mathbf{I}_k$, we see that $\mathbf{V}_k = \mathbf{X}\Gamma_k$ represents a projection of $\mathbf{X}$ onto a k-dimensional subspace of R^p. The projection onto principal coordinates is optimal out of all k-dimensional projections because it is closest to the original p-dimensional configuration. (See Theorem 14.4.1.)

This result is dual to the result in principal component analysis that the sum of the variances of the first k principal components is larger than the sum of the variances of any other k uncorrelated linear combination of the columns of $\mathbf{X}$. (See Exercise 8.2.5.)

14.4 Optimal Properties of the Classical Solution and Goodness of Fit

Given a distance matrix $\mathbf{D}$, the object of MDS is to find a configuration $\mathbf{X}$ in a low-dimensional Euclidean space R^k whose interpoint distances $\hat{d}_{rs}^2 = (\hat{\mathbf{x}}_r - \hat{\mathbf{x}}_s)'(\hat{\mathbf{x}}_r - \hat{\mathbf{x}}_s)$ say, closely match $\mathbf{D}$. The circumflex or "hat" will be used in this section to indicate that the interpoint distances $\hat{\mathbf{D}}$ for the configuration $\hat{\mathbf{X}}$ are "fitted" to the original distances $\mathbf{D}$. Similarly, let $\mathbf{B}$ denote the fitted centred inner product matrix.

Now let $\mathbf{X}$ be a configuration in R^p and let $\mathbf{L} = (\mathbf{L}_1, \mathbf{L}_2)$ be a $(p \times p)$ orthogonal matrix where $\mathbf{L}_1$ is $(p \times k)$. Then $\mathbf{XL}$ represents a projection of the configuration $\mathbf{X}$ onto the subspace of R^p spanned by the columns of $\mathbf{L}_1$. We can think of $\hat{\mathbf{X}} = \mathbf{XL}$ as a "fitted" configuration in k dimensions.

Since $\mathbf{L}$ is orthogonal, the distances between the rows of $\mathbf{X}$ are the same as the distances between the rows of $\mathbf{XL}$,

$$d_{rs}^2 = \sum_{i=1}^{p} (x_{ri} - x_{si})^2 = \sum_{i=1}^{p} (\mathbf{x}_r'\mathbf{l}_{(i)} - \mathbf{x}_s'\mathbf{l}_{(i)})^2. \tag{14.4.1}$$

If we denote the distances between the rows of $\mathbf{XL}_1$ by $\hat{D}$, then

$$\hat{d}_{rs}^2 = \sum_{i=1}^{k} (\mathbf{x}_r'\mathbf{l}_{(i)} - \mathbf{x}_s'\mathbf{l}_{(i)})^2. \tag{14.4.2}$$

Thus, $\hat{d}_{rs} \leq d_{rs}$; that is, projecting a configuration reduces the interpoint distances. Hence, a measure of the discrepancy between the original configuration $\mathbf{X}$ and the projected configuration $\hat{\mathbf{X}}$ is given by

$$\phi = \sum_{r,s=1}^{n} (d_{rs}^2 - \hat{d}_{rs}^2). \tag{14.4.3}$$

Then the classical solution to the MDS problem in k dimensions has the following optimal property:

Theorem 14.4.1 *Let $\mathbf{D}$ be a Euclidean distance matrix corresponding to a configuration $\mathbf{X}$ in R^p, and fix k $(1 \leq k < p)$. Then amongst all projections $\mathbf{XL}_1$ of $\mathbf{X}$ onto k-dimensional subspaces of R^p, the quantity (14.4.3) is minimized when $\mathbf{X}$ is projected onto its principal coordinates in k dimensions.*

Proof Using (14.4.1) and (14.4.2) we see that

$$\begin{aligned}\phi &= \sum_{r,s=1}^{n} \sum_{i=k+1}^{p} (\mathbf{x}_r'\mathbf{l}_{(i)} - \mathbf{x}_s'\mathbf{l}_{(i)})^2 \\ &= \operatorname{tr} \mathbf{L}_2' \left\{ \sum_{r,s=1}^{n} (\mathbf{x}_r - \mathbf{x}_s)(\mathbf{x}_r - \mathbf{x}_s)' \right\} \mathbf{L}_2 \\ &= 2n^2 \operatorname{tr} \mathbf{L}_2'\mathbf{S}\mathbf{L}_2\end{aligned}$$

since

$$\begin{aligned}\sum_{r,s=1}^{n} (\mathbf{x}_r - \mathbf{x}_s)(\mathbf{x}_r - \mathbf{x}_s)' &= 2n \sum_{r=1}^{n} (\mathbf{x}_r - \bar{\mathbf{x}})(\mathbf{x}_r - \bar{\mathbf{x}})' - 2 \sum_{r=1}^{n} (\mathbf{x}_r - \bar{\mathbf{x}}) \sum_{s=1}^{n} (\mathbf{x}_s - \bar{\mathbf{x}})' \\ &= 2n^2\mathbf{S}.\end{aligned}$$

Letting $\lambda_1 \geq \cdots \geq \lambda_p$ denote the eigenvalues of $n\mathbf{S}$ with standardized eigenvectors $\Gamma = (\boldsymbol{\gamma}_{(1)}, \ldots, \boldsymbol{\gamma}_{(p)})$, we can write

$$\phi = 2n \operatorname{tr} \mathbf{F}_2'\Lambda\mathbf{F}_2,$$

where $\mathbf{F}_2 = \Gamma'\mathbf{L}_2$ is a column orthonormal matrix ($\mathbf{F}_2'\mathbf{F}_2 = \mathbf{I}_{p-k}$). Using Exercise 8.2.7, we see that ϕ is minimized when $\mathbf{F}_2 = (\mathbf{0}, \mathbf{I}_{p-k})'$; that is, when $\mathbf{L}_2 = (\boldsymbol{\gamma}_{(k+1)}, \ldots, \boldsymbol{\gamma}_{(p)})$. Thus the columns of $\mathbf{L}_1$ span the space of the first k eigenvectors of $n\mathbf{S}$ and so $\mathbf{XL}_1$ represents the principal coordinates of $\mathbf{X}$ in k dimensions. Note that for this principal coordinate projection,

$$\phi = 2n(\lambda_{k+1} + \cdots + \lambda_p). \quad \blacksquare \tag{14.4.4}$$

When $\mathbf{D}$ is not necessarily Euclidean, it is more convenient to work with the matrix $\mathbf{B} = \mathbf{HAH}$. If $\hat{\mathbf{X}}$ is a fitted configuration with centred inner product matrix $\hat{\mathbf{B}}$, then a measure of the discrepancy between $\mathbf{B}$ and $\hat{\mathbf{B}}$ is given (Mardia, 1978) by

$$\psi = \sum_{r,s=1}^{n} (b_{rs} - \hat{b}_{rs})^2 = \operatorname{tr} (\mathbf{B} - \hat{\mathbf{B}})^2. \tag{14.4.5}$$

For this measure also, we can prove that the classical solution to the MDS problem is optimal.

Theorem 14.4.2 *If* $\mathbf{D}$ *is a distance matrix (not necessarily Euclidean) then for fixed* k, (14.4.5) *is minimized over all configurations* $\hat{\mathbf{X}}$ *in* k *dimensions when* $\hat{\mathbf{X}}$ *is the classical solution to the MDS problem.*

Proof Let $\lambda_1 \geqslant \cdots \geqslant \lambda_n$ denote the eigenvalues of $\mathbf{B}$, some of which might be negative, and let $\boldsymbol{\Gamma}$ denote the corresponding standardized eigenvectors. For simplicity, suppose $\lambda_k > 0$ (the situation $\lambda_k < 0$ is discussed in Exercise 14.4.2). Let $\hat{\lambda}_1 \geqslant \cdots \geqslant \hat{\lambda}_n \geqslant 0$ denote the eigenvalues of $\hat{\mathbf{B}}$. By the spectral decomposition theorem (Theorem A.6.4) we can write the symmetric matrix $\boldsymbol{\Gamma}'\hat{\mathbf{B}}\boldsymbol{\Gamma}$ as

$$\boldsymbol{\Gamma}'\hat{\mathbf{B}}\boldsymbol{\Gamma} = \mathbf{G}\hat{\boldsymbol{\Lambda}}\mathbf{G}',$$

where $\mathbf{G}$ is orthogonal. Then

$$\psi = \operatorname{tr}(\mathbf{B}-\hat{\mathbf{B}})^2 = \operatorname{tr}\boldsymbol{\Gamma}'(\mathbf{B}-\hat{\mathbf{B}})\boldsymbol{\Gamma}\boldsymbol{\Gamma}'(\mathbf{B}-\hat{\mathbf{B}})\boldsymbol{\Gamma} = \operatorname{tr}(\boldsymbol{\Lambda}-\mathbf{G}\hat{\boldsymbol{\Lambda}}\mathbf{G}')(\boldsymbol{\Lambda}-\mathbf{G}\hat{\boldsymbol{\Lambda}}\mathbf{G}').$$

We see that for fixed $\hat{\boldsymbol{\Lambda}}$ (see Exercise 14.4.2), ψ is minimized when $\mathbf{G}$ is so that

$$\psi = \sum_{i=1}^{n} (\lambda_i - \hat{\lambda}_i)^2.$$

Since $\hat{\mathbf{X}}$ lies in R^k, $\mathbf{B} = \mathbf{H}\hat{\mathbf{X}}\hat{\mathbf{X}}'\mathbf{H}$ will have at most k non-zero eigenvalues which must be non-negative. Thus, it is easy to see that ψ is minimized when

$$\hat{\lambda}_i = \begin{cases} \lambda_i, & i = 1, \ldots, k, \\ 0, & i = k+1, \ldots, n. \end{cases}$$

Hence $\hat{\mathbf{B}} = \boldsymbol{\Gamma}_1\boldsymbol{\Lambda}_1\boldsymbol{\Gamma}_1'$, where $\boldsymbol{\Gamma}_1 = (\boldsymbol{\gamma}_{(1)}, \ldots, \boldsymbol{\gamma}_{(k)})$ and $\boldsymbol{\Lambda}_1 = \operatorname{diag}(\lambda_1, \ldots, \lambda_k)$ so that $\hat{\mathbf{X}}$ can be taken to equal $\boldsymbol{\Gamma}_1\boldsymbol{\Lambda}_1^{1/2}$, the classical solution to the MDS problem in k dimensions. Note that the minimum value of ψ is given by

$$\psi = \lambda_{k+1}^2 + \cdots + \lambda_p^2. \quad \blacksquare \tag{14.4.6}$$

The above two theorems suggest possible *agreement measures* for the "proportion of a distance matrix $\mathbf{D}$ explained" by the k-dimensional classical MDS solution. Supposing $\lambda_k > 0$, these measures are (Mardia, 19

$$\alpha_{1,k} = \left(\sum_{i=1}^{k} \lambda_i \Big/ \sum_{i=1}^{n} |\lambda_i|\right) \times 100\%, \tag{14.4}$$

and

$$\alpha_{2,k} = \left(\sum_{i=1}^{k} \lambda_k^2 \Big/ \sum_{i=1}^{n} \lambda_i^2\right) \times 100\%. \tag{14.4}$$

We need to use absolute values in (14.4.7) because some of the small eigenvalues might be negative.

Example 14.4.1 We now consider the example of constructing a map of Britain from the road distances between 12 towns (Table 14.1.1). From this data, it is found that

$$\lambda_1 = 394\,473, \quad \lambda_2 = 63\,634, \quad \lambda_3 = 13\,544, \quad \lambda_4 = 10\,245,$$
$$\lambda_5 = 2465, \quad \lambda_6 = 1450, \quad \lambda_7 = 501, \quad \lambda_8 = 0,$$
$$\lambda_9 = -17, \quad \lambda_{10} = -214, \quad \lambda_{11} = -1141, \quad \lambda_{12} = -7063.$$

We note that the last four eigenvalues are negative, but they are small in relation to $\lambda_1, \ldots, \lambda_4$. We know from Theorem 14.2.1 that some negative values are expected because the distance matrix is not Euclidean.

The percentage variation explained by the first two eigenvectors is

$$\alpha_{1,2} = 92.6\% \quad \text{or} \quad \alpha_{2,2} = 99.8\%.$$

The first two eigenvectors, standardized so that $\mathbf{x}_{(i)}'\mathbf{x}_{(i)} = \lambda_i$, are

$$(45, 203, -138, 212, 189, -234, -8, -382, -32, 153, -120, 112)$$
$$(140, -18, 31, -76, 140, 31, -50, -26, -5, -27, -34, -106).$$

Since the MDS solution is invariant under rotations and translations, the coordinates have been superimposed on the true map in Figure 14.4.1 by Procrustes rotation with scaling (see Example 14.7.1). We find that the two eigenvectors closely reproduce the true map.

14.5 Seriation

14.5.1 Description

Multidimensional scaling can be used to pick out one-dimensional structure in a data set; that is, we expect the data to be parametrized by a single axis. The most common example is seriation, where we want to ascertain the chronological ordering of the data. Note that although MDS can be used to order the data in time, the *direction* of time must be determined independently.

Example 14.5.1 Consider the archeological problem of Example 8.5.1 where the similarity between graves is measured by the number of types of pottery they have in common. Using the similarity matrix $\mathbf{S}_2$ of Example 13.4.2 and the standard transformation of Theorem 14.2.2 (see Exercise 14.2.6), it is found that

$$\lambda_1 = 1.75, \quad \lambda_2 = 0.59, \quad \lambda_3 = 0.35, \quad \lambda_4 = 0.05, \quad \lambda_5 = \lambda_6 = 0,$$

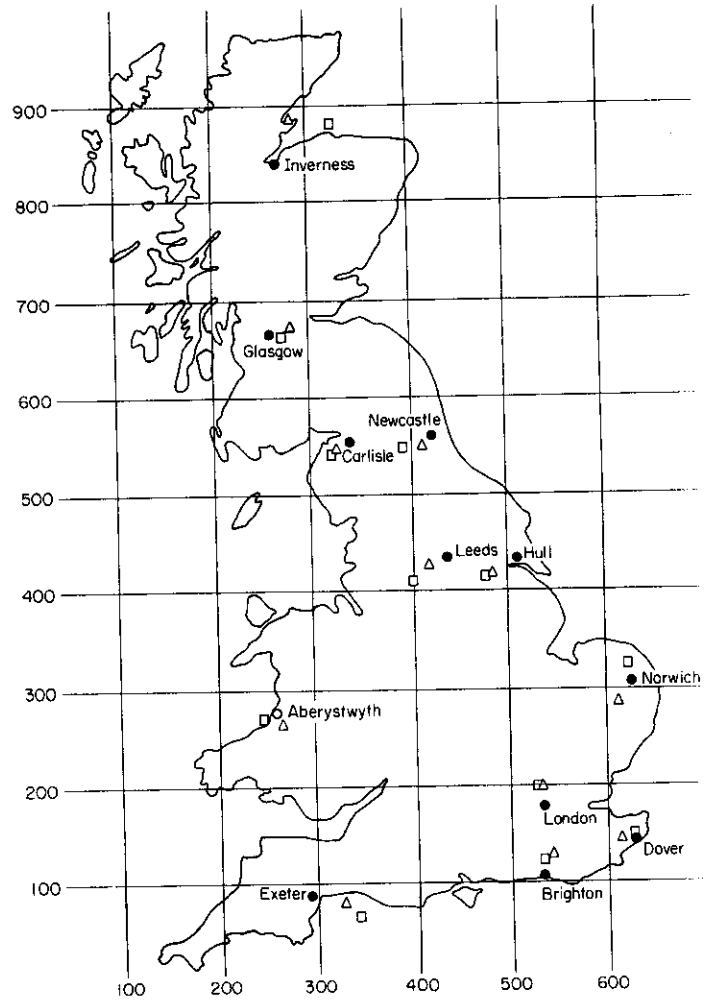


Figure 14.4.1 MDS solutions for the road data in Table 1.2.2. ●, *original points.* △, *classical solution;* □, *Shepard–Kruskal solution.*

with coordinates in two dimensions

$$(-0.60, 0.77, -0.19, -0.60, 0.64, -0.01),$$

and

$$(-0.15, 0.20, 0.60, -0.15, -0.35, -0.14).$$

See Figure 14.5.1. The coordinates in one dimension suggest the order (A, D), C, F, E, B, which is similar but not identical with the ordering given by correspondence analysis (Example 8.5.1).

It is often a good idea to plot the data in more than one dimension to see if the data is in fact one-dimensional. For example, the artificial data in the above example does not particularly seem to lie in one dimension. However, even when the data is truly one-dimensional, it need not lie on the axis of the first dimension but can sometimes lie on a curve as the following section shows.

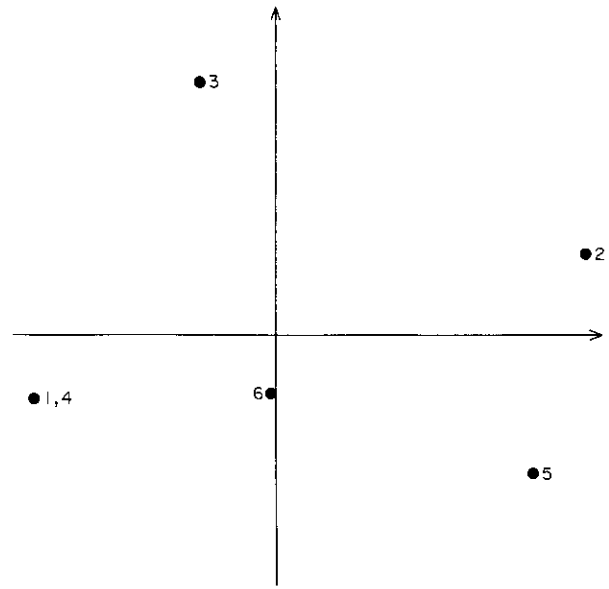


Figure 14.5.1 Classical MDS solution in two dimensions using similarity matrix $\mathbf{S}_2$ *for grave data in Example 14.5.1.*

14.5.2 Horseshoe effect

In some situations we can measure accurately the distance between two objects when they are close together but not when they are far apart. Distances which are "moderate" and those which are "large" appear to be the same. For example, consider the archeology example described above. Graves which are close together in time will have some varieties of pottery in common, but those which are separated by more than a certain gap of time will have *no* varieties in common.

This merging of all "large" distances tends to pull the farthest objects closer together and has been labelled by D. G. Kendall (1971) the "horseshoe effect". This effect can be observed clearly in the following artificial example.

Example 14.5.2 (D. G. Kendall, 1971) Consider the (51×51) similarity matrix $\mathbf{C}$ defined by

$$c_{rr}=9, \qquad c_{rs}=\begin{cases} 8 & \text{if } 1\leqslant |r-s|\leqslant 3, \\ \cdots & \\ 1 & \text{if } 22\leqslant |r-s|\leqslant 24, \\ 0 & \text{if } |r-s|\geqslant 25. \end{cases}$$

Using the standard transformation from similarities to distances leads to eight negative eigenvalues (varying from -0.09 to -2.07) and 43 non-negative eigenvalues,

$$126.09, 65.94, 18.17, 7.82, 7.61, 7.38, 7.02, 5.28, 3.44, \ldots, 0.10, 0.$$

A plot of the configuration in two dimensions is given in Figure 14.5.2.

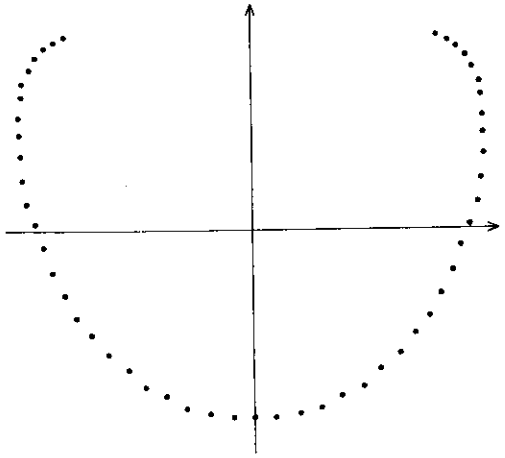


Figure 14.5.2 Two-dimensional representation of Kendall's matrix.

The furthest points are pulled together so that the configuration looks roughly like part of the circumference of a circle. Note that while the ordering is clear from this figure, it is not ascertainable from the one-dimensional classical solution.

14.6 Non-metric Methods

Implicit in the preceding sections is the assumption that there is a "true" configuration in k dimensions with interpoint distances δ_{rs}. We wish to reconstruct this configuration using an observed distance matrix $\mathbf{D}$ whose elements are of the form

$$d_{rs}=\delta_{rs}+e_{rs}. \tag{14.6.1}$$

Here the e_{rs} represent errors of measurement plus distortion errors arising because the distances do not exactly correspond to a configuration in R^k.

However, in some situations it is more realistic to hypothesize a less rigid relationship between d_{rs} and δ_{rs}; namely, suppose

$$d_{rs}=f(\delta_{rs}+e_{rs}), \tag{14.6.2}$$

where f is an unknown monotone increasing function. For this "model", the only information we can use to reconstruct the δ_{rs} is the *rank order* of the d_{rs}. For example, for the road map data, we could try to reconstruct the map of England using the information

the quickest journey is that from Brighton to London;

the next quickest journey is that from Hull to Leeds;

...

the longest journey is that from Dover to Inverness.

In this non-metric approach $\mathbf{D}$ is not thought of as a "distance" matrix but as a "dissimilarity" matrix. In fact the non-metric approach is often most appropriate when the data is presented as a similarity matrix. For in this situation the transformation from similarities to distances is somewhat arbitrary and perhaps the strongest statement one should make is that greater similarity implies less dissimilarity.

An algorithm to construct a configuration based on the rank order information has been developed by Shepard (1962*a*,*b*) and Kruskal (1964).

Shepard–Kruskal algorithm

(a) Given a dissimilarity matrix $\mathbf{D}$, order the off-diagonal elements so that

$$d_{r_1 s_1} \leqslant \cdots \leqslant d_{r_m s_m}, \qquad m = \frac{1}{2}n(n-1), \tag{14.6.}$$

where $(r_1, s_1), \ldots, (r_m, s_m)$ denote all pairs of unequal subscripts $r_i < s_i$. Say that numbers d_{rs}^* are monotonically related to the d_{rs} (and write $d_{rs}^* \overset{\text{mon}}{\sim} d_{rs}$) if

$$d_{rs} < d_{uv} \Rightarrow d_{rs}^* \leqslant d_{uv}^* \qquad \text{for all } r<s,\ u<v. \tag{14.6.4}$$

(b) Let $\hat{\mathbf{X}}(n \times k)$ be a configuration in R^k with interpoint distances $\hat{d}_{rs}$. Define the (squared) *stress* of $\hat{\mathbf{X}}$ by

$$S^2(\hat{\mathbf{X}}) = \min \sum_{r<s} (d_{rs}^* - \hat{d}_{rs})^2 \Big/ \sum_{r<s} \hat{d}_{rs}^2, \tag{14.6.}$$

where the minimum is taken over d_{rs}^* such that $d_{rs}^* \overset{\text{mon}}{\sim} d_{rs}$. The d_{rs}^* which minimizes (14.6.5) represent the *least squares monotone regression* of $\hat{d}_{rs}$ on d_{rs}. Thus (14.6.5) represents the extent to which the rank order of the $\hat{d}_{rs}$ disagrees with the rank order of the d_{rs}. If the rank orders match exactly (which is very rare in practice), then $S(\hat{\mathbf{X}}) = 0$. The presence of the denominator in (14.6.5) standardizes the stress and makes it invariant under transformations of the form $\mathbf{y}_r = c\mathbf{x}_r$, $r = 1, \ldots, n$, $c \neq 0$. The stress is also invariant under transformations of the form $\mathbf{y}_r = \mathbf{A}\mathbf{x}_r + \mathbf{b}$ when $\mathbf{A}$ is orthogonal.

(c) For each dimension k, the configuration which has the smallest stress is called the *best fitting configuration in k dimensions*. Let

$$S_k = \min_{\hat{\mathbf{X}}(n \times k)} S(\hat{\mathbf{X}})$$

denote this minimal stress.

(d) To choose the correct dimension, calculate $S_1, S_2, \ldots,$ until the value becomes low. Say, for example, S_k is low for $k = k_0$. Since S_k is a decreasing function of k, $k = k_0$ is the "right dimension". A rule of thumb is provided by Kruskal to judge the tolerability of S_k: $S_k \geqslant 20\%$, poor; $S_k = 10\%$, fair; $S_k \leqslant 5\%$, good; $S_k = 0$, perfect.

Remarks (1) The "best configuration" starting from an arbitrary initial configuration can be obtained by using a computer routine developed by

Kruskal (1964) which utilizes the method of steepest descent to find the local minimum. The initial configuration can be taken as the classical solution. Unfortunately, there is no way of distinguishing in practice between a local minimum and the global minimum.

(2) The Shepard–Kruskal solution is invariant under rotation, translation, and uniform expansion or contraction of the best-fitting configuration.

(3) The Shepard–Kruskal solution is non-metric since it utilizes only the rank orders (14.6.3). However, we still need a sufficiently objective numerical measure of distance to determine the rank order of the d_{rs}.

(4) *Similarities* The non-metric method works just as well with similarities as with dissimilarities. One simply changes the direction of the inequalities.

(5) *Missing values* The Shepard–Kruskal method is easily adapted to the situation where there are missing values. One simply omits the missing dissimilarities in the ordering (14.6.3) and deletes the corresponding terms from the numerator and denominator of (14.6.5). As long as not too many values are missing, the method still seems to work well.

(6) *Treatment of ties* The constraint given by (14.6.4) is called the *primary treatment of ties* (PTT). If $d_{rs} = d_{uv}$ then no constraint is made on d_{rs}^* and d_{uv}^*. An alternative constraint, called the *secondary treatment of ties* (STT) is given by

$$d_{rs} \leqslant d_{uv} \Rightarrow d_{rs}^* \leqslant d_{uv}^*,$$

which has the property that $d_{rs} = d_{uv} \Rightarrow d_{rs}^* = d_{uv}^*$. However, one must be cautious when using STT on data with a large number of ties. The use of STT on data such as Example 14.5.2 leads to the horseshoe effect (see, D. G. Kendall, 1971).

(7) *Comparison of methods* The computation is simpler for the classical method than it is for the non-metric method. It is not known how robust the classical method is to monotone transformations of the distance function; however, both methods seem to give similar answers when applied to well-known examples in the field. Figure 14.4.1 gives the two solutions for the road data. For the Shepard–Kruskal solution for the Morse code data, see Shepard (1963).

(8) We have not commented on the non-metric method of Guttman (1968). For a mathematical and empirical analysis of multidimensional scaling algorithms of Kruskal's M-D-SCA1 and Guttman–Lingoes' SAA-I, we refer to Lingoes and Roskam (1973) which also contains certain recommendations for improvement of these algorithms.

14.7 Goodness of Fit Measure: Procrustes Rotation

We now describe a *goodness of fit* measure (Green, 1952; Gower, 1971*b*), used to compare two configurations. Let $\mathbf{X}$ be the $(n \times p)$ matrix of the coordinates of n points obtained from $\mathbf{D}$ by one technique. Suppose that $\mathbf{Y}$ is the $(n \times q)$ matrix of coordinates of another set of points obtained by another technique, or using another measure of distance. Let $q \leq p$. By adding columns of zeros to $\mathbf{Y}$, we may also assume $\mathbf{Y}$ to be $(n \times p)$.

The measure of goodness of fit adopted is obtained by moving the points $\mathbf{y}_r$ relative to the points $\mathbf{x}_r$ until the "residual" sum of squares

$$\sum_{r=1}^{n} (\mathbf{x}_r - \mathbf{y}_r)'(\mathbf{x}_r - \mathbf{y}_r) \tag{14.7.1}$$

is minimal. We can move $\mathbf{y}_r$ relative to $\mathbf{x}_r$ through rotation, reflection, and translation, i.e. by

$$\mathbf{A}'\mathbf{y}_r + \mathbf{b}, \qquad r = 1, \ldots, n, \tag{14.7.2}$$

where $\mathbf{A}'$ is a $(p \times p)$ orthogonal matrix. Hence, we wish to solve

$$R^2 = \min_{\mathbf{A},\mathbf{b}} \sum_{r=1}^{n} (\mathbf{x}_r - \mathbf{A}'\mathbf{y}_r - \mathbf{b})'(\mathbf{x}_r - \mathbf{A}'\mathbf{y}_r - \mathbf{b}) \tag{14.7.3}$$

for $\mathbf{A}$ and $\mathbf{b}$. Note that $\mathbf{A}$ and $\mathbf{b}$ are found by least squares. Their values are given in the following theorem.

Theorem 14.7.1 *Let $\mathbf{X}(n \times p)$ and $\mathbf{Y}(n \times p)$ be two configurations of n points, for convenience centred at the origin, so $\bar{\mathbf{x}} = \bar{\mathbf{y}} = \mathbf{0}$. Let $\mathbf{Z} = \mathbf{Y}'\mathbf{X}$ and using the singular value decomposition theorem (Theorem A.6.5), write*

$$\mathbf{Z} = \mathbf{V}\boldsymbol{\Gamma}\mathbf{U}', \tag{14.7.4}$$

where $\mathbf{V}$ and $\mathbf{U}$ are orthogonal $(p \times p)$ matrices and $\boldsymbol{\Gamma}$ is a diagonal matrix of non-negative elements. Then the minimizing values of $\mathbf{A}$ and $\mathbf{b}$ in (14.7.3) are given by

$$\hat{\mathbf{b}} = \mathbf{0}, \qquad \hat{\mathbf{A}} = \mathbf{V}\mathbf{U}', \tag{14.7.5}$$

and further

$$R^2 = \operatorname{tr}\mathbf{X}\mathbf{X}' + \operatorname{tr}\mathbf{Y}\mathbf{Y}' - 2\operatorname{tr}\boldsymbol{\Gamma}. \tag{14.7.6}$$

Proof On differentiating with respect to $\mathbf{b}$, we have

$$\hat{\mathbf{b}} = \bar{\mathbf{x}} - \mathbf{A}'\bar{\mathbf{y}} \tag{14.7.7}$$

where $\bar{\mathbf{y}} = \sum \mathbf{y}_r/n$, $\bar{\mathbf{x}} = \sum \mathbf{x}_r/n$. Since both configurations are centred, $\hat{\mathbf{b}} = \mathbf{0}$

Then we can rewrite (14.7.3) as

$$R^2 = \min_{\mathbf{A}} \operatorname{tr}(\mathbf{X} - \mathbf{Y}\mathbf{A})(\mathbf{X} - \mathbf{Y}\mathbf{A})' = \operatorname{tr}\mathbf{X}\mathbf{X}' + \operatorname{tr}\mathbf{Y}\mathbf{Y}' - 2\max_{\mathbf{A}} \operatorname{tr}\mathbf{X}'\mathbf{Y}\mathbf{A}. \tag{14.7.8}$$

The constraints on $\mathbf{A}$ are $\mathbf{A}\mathbf{A}' = \mathbf{I}$, i.e. $\mathbf{a}_i'\mathbf{a}_i = 1$, $\mathbf{a}_i'\mathbf{a}_j = 0$, $i \neq j$, where $\mathbf{a}_i'$ is the ith row of $\mathbf{A}$. Hence there are $p(p+1)/2$ constraints.

Let $\frac{1}{2}\boldsymbol{\Lambda}$ be a $(p \times p)$ symmetric matrix of Lagrange multipliers for these constraints. The aim is to maximize

$$\operatorname{tr}\{\mathbf{Z}'\mathbf{A} - \tfrac{1}{2}\boldsymbol{\Lambda}(\mathbf{A}\mathbf{A}' - \mathbf{I})\}, \tag{14.7.9}$$

where $\mathbf{Z}' = \mathbf{X}'\mathbf{Y}$. By direct differentiation it can be shown that

$$\frac{\partial}{\partial \mathbf{A}} \operatorname{tr}(\mathbf{Z}'\mathbf{A}) = \mathbf{Z}, \qquad \frac{\partial}{\partial \mathbf{A}} \operatorname{tr}(\boldsymbol{\Lambda}\mathbf{A}\mathbf{A}') = 2\boldsymbol{\Lambda}\mathbf{A}. \tag{14.7.10}$$

Hence on differentiating (14.7.9) and equating the derivatives to zero, we find that $\mathbf{A}$ must satisfy

$$\mathbf{Z} = \boldsymbol{\Lambda}\mathbf{A}. \tag{14.7.11}$$

Write $\mathbf{Z}$ using (14.7.4). Noting that $\boldsymbol{\Lambda}$ is symmetric and that $\mathbf{A}$ is to be orthogonal, we get, from (14.7.11),

$$\boldsymbol{\Lambda}^2 = \mathbf{Z}\mathbf{A}'\mathbf{A}\mathbf{Z} = \mathbf{Z}\mathbf{Z}' = (\mathbf{V}\boldsymbol{\Gamma}\mathbf{U}')(\mathbf{U}\boldsymbol{\Gamma}\mathbf{V}').$$

Thus we can take $\boldsymbol{\Lambda} = \mathbf{V}\boldsymbol{\Gamma}\mathbf{V}'$. Substituting this value of $\boldsymbol{\Lambda}$ in (14.7.11) we see that

$$\hat{\mathbf{A}} = \mathbf{V}\mathbf{U}' \tag{14.7.12}$$

is a solution of (14.7.11). Note that $\hat{\mathbf{A}}$ is orthogonal. Using this value of $\hat{\mathbf{A}}$ in (14.7.8) gives (14.7.6).

Finally, to verify that $\hat{\mathbf{A}}$ maximizes (14.7.9) (and is not just a stationary point) we must differentiate (14.7.9) with respect to $\mathbf{A}$ a second time. For this purpose it is convenient to write $\mathbf{A}$ as a vector $\mathbf{a} = (\mathbf{a}_{(1)}', \ldots, \mathbf{a}_{(p)}')'$. Then (14.7.9) is a quadratic function of the elements of $\mathbf{a}$ and the second derivative of (14.7.9) with respect to $\mathbf{a}$ can be expressed as the matrix $-\mathbf{I}_p \otimes \boldsymbol{\Lambda}$. Since $\boldsymbol{\Lambda} = \mathbf{V}\boldsymbol{\Gamma}\mathbf{V}'$, and the diagonal elements of $\boldsymbol{\Gamma}$ are non-negative, we see that the second derivative matrix is negative semi-definite. Hence $\hat{\mathbf{A}}$ maximizes (14.7.9). ■

We have assumed that the column means of $\mathbf{X}$ and $\mathbf{Y}$ are zero. Then the "best" rotation of $\mathbf{Y}$ relative to $\mathbf{X}$ is $\mathbf{Y}\hat{\mathbf{A}}$, where $\hat{\mathbf{A}}$ is given by (14.7.12), and $\hat{\mathbf{A}}$ is called the *Procrustes rotation* of $\mathbf{Y}$ relative to $\mathbf{X}$. Noting from (14.7.4) that $\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X} = \mathbf{U}\boldsymbol{\Gamma}^2\mathbf{U}'$, we can rewrite (14.7.8) as

$$R^2 = \operatorname{tr}\mathbf{X}\mathbf{X}' + \operatorname{tr}\mathbf{Y}\mathbf{Y}' - 2\operatorname{tr}(\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X})^{1/2}. \tag{14.7.12}$$

It can be seen that (14.7.8) is zero if and only if the $\mathbf{y}_r$ can be rotated to the $\mathbf{x}_r$ exactly.

Scale factor If the scales of two configurations are different, then the transformation (14.7.2) should be of the form

$$c\mathbf{A}'\mathbf{y}_r + \mathbf{b},$$

where $c > 0$. Following the above procedure, it can be seen that

$$\hat{c} = (\operatorname{tr} \boldsymbol{\Gamma})/(\operatorname{tr} \mathbf{YY}') \tag{14.7.14}$$

and the other estimates remain as before. This transformation is called the *Procrustes rotation with scaling* of $\mathbf{Y}$ relative to $\mathbf{X}$. Then the minimum residual sum of squares is given by

$$R^2 = \operatorname{tr}(\mathbf{XX}') + \hat{c}^2 \operatorname{tr}(\mathbf{YY}') - 2\hat{c} \operatorname{tr}(\mathbf{X}'\mathbf{YY}'\mathbf{X})^{1/2}, \tag{14.7.15}$$

where $\hat{c}$ is given by (14.7.14). Note that this procedure is not symmetric with respect to $\mathbf{X}$ and $\mathbf{Y}$. Symmetry can be obtained by selecting scales so that

$$\operatorname{tr}(\mathbf{XX}') = \operatorname{tr}(\mathbf{YY}').$$

For an excellent review of this topic, see Sibson (1978).

Example 14.7.1 The actual Ordnance Survey coordinates of the towns in Table 14.1.1 are

	1	2	3	4	5	6	7	8	9	10	11
E	257	529	339	629	292	259	508	265	433	533	420
N	279	104	554	142	90	665	433	842	438	183	563

Treating these quantities as planar coordinates $\mathbf{X}(12\times 2)$ (the curvature of the Earth has little effect), and the first two eigenvectors from the classical MDS solution for the data given in Example 14.4.1 as $\mathbf{Y}$, it is found that

$$\mathbf{X}'\mathbf{Y} = \begin{pmatrix} 182\,119.068 & -91\,647.926 \\ -495\,108.159 & -25\,629.185 \end{pmatrix}, \quad \mathbf{U} = \begin{pmatrix} -0.347\,729 & -0.937\,\cdots \\ 0.937\,595 & -0.34\,\cdots \end{pmatrix}$$

$$\boldsymbol{\Gamma} = \begin{pmatrix} 527\,597.29 & 0 \\ 0 & 94\,851.13 \end{pmatrix}, \quad \mathbf{V} = \begin{pmatrix} -0.999\,890 & 0.014\,858 \\ 0.014\,858 & 0.999\,890 \end{pmatrix}$$

This leads to the transforming of the $\mathbf{y}$s to match the $\mathbf{x}$s by

$$\mathbf{y}_r^* = c\mathbf{A}'\mathbf{y}_r + \mathbf{b},$$

where

$$c = 1.358\,740, \quad \mathbf{A} = \mathbf{VU}' = \begin{pmatrix} 0.333\,760, & -0.942\,658 \\ -0.942\,658, & -0.333\,760 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 424.250 \\ 383.417 \end{pmatrix}.$$

The transformation has been used on $\mathbf{y}_r$ to obtain $\mathbf{y}_r^*$, and these $\mathbf{y}_r^*$ are plotted in Figure 14.4.1 together with the $\mathbf{x}_r$. We have

$$\operatorname{tr}\mathbf{XX}' = 853\,917, \qquad \operatorname{tr}\mathbf{YY}' = 458\,107.$$

Hence from (14.7.15), the residual is $R^2 = 8172$.

The Shepard–Kruskal solution has a stress of 0.0404. Using this solution as $\mathbf{Y}$ and the Ordnance Survey coordinates again as $\mathbf{X}$ leads to a residual of $R^2 = 13\,749$. Hence the classical solution fits a bit better for this data. Of course Figure 14.4.1 shows little difference between the two solutions.

If we are given two distance matrices $\mathbf{D}_1$ and $\mathbf{D}_2$ *but* not the corresponding points, (14.7.6) cannot be computed without using some method to compute the "points". The first two terms are expressible in terms of $\mathbf{D}_1$ and $\mathbf{D}_2$ but not $\operatorname{tr}\boldsymbol{\Gamma}$.

*14.8 Multi-sample Problem and Canonical Variates

Consider the case of g p-variate populations with means $\boldsymbol{\mu}_r$, $r = 1, \ldots, g$, and common covariance matrix $\boldsymbol{\Sigma}$. If we are given a data matrix $\mathbf{Y}$ representing samples of size n_r from the rth group, $r = 1, \ldots, g$, let $\bar{\mathbf{y}}_r$ denote the rth sample mean, and estimate the common covariance matrix $\boldsymbol{\Sigma}$ by $\mathbf{W}/\nu$, where $\mathbf{W}$ is the within-groups sum of squares and products (SSP) matrix with ν degrees of freedom.

Assume that the overall (unweighted) mean $\bar{\mathbf{y}}$ is

$$\bar{\mathbf{y}} = \sum_{r=1}^{n} \bar{\mathbf{y}}_r = \mathbf{0}. \tag{14.8.1}$$

We shall work with the Mahalanobis distances

$$d_{rs}^2 = \nu(\bar{\mathbf{y}}_r - \bar{\mathbf{y}}_s)'\mathbf{W}^{-1}(\bar{\mathbf{y}}_r - \bar{\mathbf{y}}_s). \tag{14.8.2}$$

It is easily checked that if $\mathbf{B}$ is defined by (14.2.3), then

$$\mathbf{B} = \nu\bar{\mathbf{Y}}\mathbf{W}^{-1}\bar{\mathbf{Y}}',$$

where $\bar{\mathbf{Y}}' = (\bar{\mathbf{y}}_1, \ldots, \bar{\mathbf{y}}_g)$.

Thus $\mathbf{B} \geqslant 0$ and so $\mathbf{D}$ is Euclidean. Let $\mathbf{X}$ be the configuration for $\mathbf{B}$ defined in Theorem 14.2.1, and fix k, $1 \leqslant k \leqslant p$. Then, the first k columns of $\mathbf{X}$ can be regarded as the coordinates of points representing the g means in k dimensions ($k \leqslant p$). This configuration has the optimal property that it is the "best" representation in k dimensions.

Note that $\bar{\mathbf{Y}}'\bar{\mathbf{Y}}$ is the (unweighted) between-groups SSP matrix. Let $\mathbf{l}_i$ denote the ith canonical vector of Section 12.5 using this unweighted between-groups SSP matrix; that is, define $\mathbf{l}_i$ by

$$\nu \mathbf{W}^{-1}\bar{\mathbf{Y}}'\bar{\mathbf{Y}}\mathbf{l}_i = \lambda_i \mathbf{l}_i, \qquad \nu^{-1}\mathbf{l}_i'\mathbf{W}\mathbf{l}_i = 1, \qquad i = 1, \ldots, \min(p, g)$$

where λ_i is the ith eigenvalue of $\nu\mathbf{W}^{-1}\bar{\mathbf{Y}}'\bar{\mathbf{Y}}$, which is the same as the ith eigenvalue of $\mathbf{B}$. Then the scores of the g groups on the ith canonical coordinate are given by

$$\bar{\mathbf{Y}}\mathbf{l}_i.$$

Since $\mathbf{B}\bar{\mathbf{Y}}\mathbf{l}_i = \lambda_i \bar{\mathbf{Y}}\mathbf{l}_i$ and $\mathbf{l}_i'\bar{\mathbf{Y}}'\bar{\mathbf{Y}}\mathbf{l}_i = \lambda_i$, we see from (14.2.5) that

$$\bar{\mathbf{Y}}\mathbf{l}_i = \mathbf{x}_{(i)},$$

so that $\bar{\mathbf{Y}}\mathbf{l}_i$ is also the ith eigenvector of $\mathbf{B}$. Thus, the canonical means in k dimensions, that is, the scores of the first k canonical variates on the g groups, are the same as the coordinates given by Theorem 14.2.1.

Exercises and Complements

14.2.1 Using (14.2.6), show that (14.2.9) can be written in the form (14.2.10).

14.2.2 In the notation of Theorem 14.2.1 show that

(a) $b_{rr} = \bar{a}_{\cdot\cdot} - 2\bar{a}_{r\cdot}$, $b_{rs} = a_{rs} - \bar{a}_{r\cdot} - \bar{a}_{\cdot s} + \bar{a}_{\cdot\cdot}$, $r \neq s$;

(b) $\mathbf{B} = \sum_{i=1}^{p} \mathbf{x}_{(i)}\mathbf{x}_{(i)}'$;

(c) $\sum_{r=1}^{n} \lambda_r = \sum_{r=1}^{n} b_{rr} = \frac{1}{2n} \sum_{r,s=1}^{n} d_{rs}^2$.

14.2.3 (Gower, 1968) Let $\mathbf{D} = (d_{rs})$ be an $(n \times n)$ Euclidean distance matrix with configuration $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ in p-dimensional principal coordinates, given by Theorem 14.2.1. Suppose we wish to add an additional point to the configuration using distances $d_{r,n+1}$, $r = 1, \ldots, n$ (which we know to be Euclidean), allowing for a $(p+1)$th dimension. If the first n points are represented by $(x_{r1}, \ldots, x_{rp}, 0)'$, $r = 1, \ldots, n$, then show that

the $(n+1)$th point is given by

$$\mathbf{x}_{n+1} = (x_{n+1,1}, \ldots, x_{n+1,p}, x_{n+1,p+1})' = (\mathbf{x}', y)', \text{ say}$$

where

$$\mathbf{x} = \tfrac{1}{2}\mathbf{\Lambda}^{-1}\mathbf{X}'\mathbf{f}, \qquad \mathbf{f} = (f_1, \ldots, f_n)', \qquad f_r = b_{rr} - d_{r,n+1}^2,$$

and

$$y^2 = \frac{1}{n}\sum_{r=1}^{n} d_{r,n+1}^2 - \frac{1}{n}\sum_{r=1}^{n} b_{rr} - \mathbf{x}'\mathbf{x}.$$

Hence $\mathbf{x}$ is uniquely determined but y is determined only in value, not in sign. Give the reason. (Hint: substitute $f_r = 2\mathbf{x}_r'\mathbf{x}_{n+1} - \mathbf{x}_{n+1}'\mathbf{x}_{n+1}$ for f_r in terms of $\mathbf{x}_1, \ldots, \mathbf{x}_{n+1}$ to verify the formula for $\mathbf{x}$.)

14.2.4 If $\mathbf{C}$ is p.s.d. then show that $c_{ii} + c_{jj} - 2c_{ij} > 0$. Show that the distance d_{ij} defined by $d_{ij}^2 = c_{ii} + c_{jj} - 2c_{ij}$ satisfies the triangle inequality.

14.2.5 For the Bhattacharyya distance matrix $\mathbf{D}$ given in Example 13.4.1, the eigenvalues of $\mathbf{B}$ are

$$\lambda_1 = 318.97, \qquad \lambda_2 = 167.72, \qquad \lambda_3 = 11.11, \qquad \lambda_4 = 0.$$

Hence, $\mathbf{D}$ is Euclidean and the two-dimensional representation accounts for $\alpha_{1,2} = 98\%$ of the variation.

Show that the principal coordinates in two dimensions for Eskimo, Bantu, English, and Korean are, respectively,

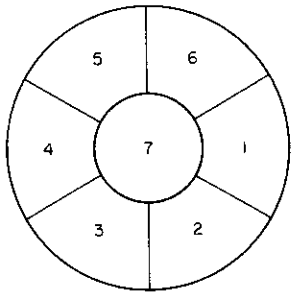
$$(9.69, 7.29), \qquad (-11.39, -2.51), \qquad (-6.00, 4.57), \qquad (7.70, -9.34).$$

Plot these points and comment on the conclusions drawn in Example 13.4.1.

14.2.6 For the grave data (Example 13.4.2) using the similarity matrix $\mathbf{S}_2$ show that the distance matrix given by the standard transformation is

$$\mathbf{D} = \begin{bmatrix} 0 & \sqrt{10} & 2 & 0 & \sqrt{8} & 2 \\ & 0 & \sqrt{6} & \sqrt{10} & \sqrt{2} & \sqrt{6} \\ & & 0 & 2 & \sqrt{8} & 2 \\ & & & 0 & \sqrt{8} & 2 \\ & & & & 0 & 2 \\ & & & & & 0 \end{bmatrix}.$$

14.2.7 Suppose that $1, 2, \ldots, 7$ are regions (enclosed by unbroken lines) in a country arranged as in Exercise Figure 1. Let the distance matrix be



Exercise Figure 1 Seven regions in a country.

constructed by counting the minimum number of boundaries crossed to pass from region i to region j.

Show that the distance matrix is given by

$$\begin{bmatrix} 0 & 1 & 2 & 2 & 2 & 1 & 1 \\ & 0 & 1 & 2 & 2 & 2 & 1 \\ & & 0 & 1 & 2 & 2 & 1 \\ & & & 0 & 1 & 2 & 1 \\ & & & & 0 & 1 & 1 \\ & & & & & 0 & 1 \\ & & & & & & 0 \end{bmatrix}$$

Show that the distances constructed in this way obey the triangle inequality $d_{ik} \leqslant d_{ij} + d_{jk}$, but by showing that the eigenvalues of the matrix $\mathbf{B}$ are

$$\lambda_1 = \lambda_2 = \tfrac{7}{2}, \quad \lambda_3 = \lambda_4 = \tfrac{1}{2}, \quad \lambda_5 = 0, \quad \lambda_6 = -\tfrac{1}{7}, \quad \lambda_7 = -1,$$

deduce that this metric is non-Euclidean.

Since $\lambda_1 = \lambda_2$, select any two orthogonal eigenvectors corresponding to λ_1 and λ_2 and, by plotting the seven points so obtained, show that the original map is reconstructed. As in Example 14.2.1, the points are vertices of a hexagon with centre at the origin.

14.2.8 (Lingoes, 1971; Mardia, 1978) Let $\mathbf{D}$ be a distance matrix. Show that for some real number a, there exists a Euclidean configuration in $p \leqslant n-2$ dimensions with interpoint distances d^*_{rs} satisfying

$$d^{*2}_{rs} = d^2_{rs} - 2a, \quad r \neq s; \qquad d^*_{rr} = 0.$$

Thus d^{*2}_{rs} is a linear function of d^2_{rs}, so the configuration preserves the

rank order of the distances. (Hint: show that the matrix $\mathbf{D}^*$ leads to $\mathbf{A}^*$ and $\mathbf{B}^*$ given by

$$\mathbf{A}^* = (-\tfrac{1}{2}d^{*2}_{rs}) = \mathbf{A} - a(\mathbf{I}-\mathbf{J}), \qquad \mathbf{B}^* = \mathbf{H}\mathbf{A}^*\mathbf{H} = \mathbf{B} - a\mathbf{H}.$$

If $\mathbf{B}$ has eigenvalues $\lambda_1 \geqslant \cdots \geqslant \lambda_u > 0 \geqslant \lambda'_1 \geqslant \cdots \geqslant \lambda'_v$, then $\mathbf{B}^*$ has eigenvalues $\lambda_r - a$, $r = 1, \ldots, u$; 0; and $\lambda'_r - a$, $r = 1, \ldots, v$. Then the choice $a = \lambda'_v$ makes $\mathbf{B}^*$ p.s.d. of rank at most $n-2$.)

14.4.1 Let $\mathbf{l}_1, \ldots, \mathbf{l}_p$ be orthonormal vectors in $R^q (p \leqslant q)$ and let $\mathbf{z}_r$, $r = 1, \ldots, n$, be points in R^q. Let H_r denote the foot of the perpendicular of $\mathbf{z}_r$ on the subspace spanned by $\mathbf{l}_1, \ldots, \mathbf{l}_p$. Show that with respect to the new coordinate system with axes $\mathbf{l}_1, \ldots, \mathbf{l}_p$, the coordinates of H_r are $(\mathbf{l}'_1\mathbf{z}_r, \ldots, \mathbf{l}'_p\mathbf{z}_r)'$. What modification must be made if the $\mathbf{l}_i$ are orthogonal but not orthonormal?

14.4.2 Let $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_p)$, where $\lambda_1 \geqslant \cdots \geqslant \lambda_p$ are real numbers, and let $\hat{\mathbf{\Lambda}} = \operatorname{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_p)$, where $\hat{\lambda}_1 \geqslant \cdots \geqslant \hat{\lambda}_p \geqslant 0$ are non-negative numbers. Show that minimizing

$$\operatorname{tr}(\mathbf{\Lambda} - \mathbf{G}\hat{\mathbf{\Lambda}}\mathbf{G}')^2$$

over orthogonal matrices $\mathbf{G}$ is equivalent to maximizing

$$\operatorname{tr}(\mathbf{\Lambda}\mathbf{G}\hat{\mathbf{\Lambda}}\mathbf{G}') = \sum_{i,j=1}^{p} \lambda_i \hat{\lambda}_j g^2_{ij} = \sum_{i=1}^{p} \lambda_i h_i = \phi(\mathbf{h}) \quad \text{say,}$$

where

$$h_i = \sum_{j=1}^{p} \hat{\lambda}_j g^2_{ij} \geqslant 0 \quad \text{and} \quad \sum_{i=1}^{p} h_i = \sum_{j=1}^{p} \hat{\lambda}_j.$$

Show that $\phi(\mathbf{h})$ is maximized over such vectors $\mathbf{h}$ when $h_i = \hat{\lambda}_i$ for $i = 1, \ldots, p$; that is, when $\mathbf{G} = \mathbf{I}$.