Introduction to Applied Multivariate Analysis with

Multidimensional Scaling

Classical MD Scaling

Takes an input matrix giving dissimilarities between pairs of items and outputs a coordinate matrix.

To begin our account of the method, we shall assume that the proximity matrix we are dealing with is a matrix of Euclidean distances, $\mathbf{D}$, derived from a raw $n \times q$ data matrix, $\mathbf{X}$. In Chapter 1, we saw how to calculate Euclidean distances from $\mathbf{X}$; classical multidimensional scaling is essentially concerned with the reverse problem: given the distances, how do we find $\mathbf{X}$?

Classical MD Scaling

Want to find b's in terms of the d's to derive the coordinate values by factoring B. Need to introduce location constraint.

First assume $\mathbf{X}$ is known and consider the $n \times n$ *inner products matrix*, $\mathbf{B}$

$$\mathbf{B} = \mathbf{X}\mathbf{X}^\top. \tag{4.1}$$

The elements of $\mathbf{B}$ are given by

$$b_{ij} = \sum_{k=1}^{q} x_{ik} x_{jk}. \tag{4.2}$$

It is easy to see that the squared Euclidean distances between the rows of $\mathbf{X}$ can be written in terms of the elements of $\mathbf{B}$ as

$$d_{ij}^2 = b_{ii} + b_{jj} - 2b_{ij}. \tag{4.3}$$

Classical MD Scaling

Measure of fit is how well the proximities match distances in the geometrical model.

Using all q-dimensions will lead to complete recovery of the original Euclidean distance matrix. The best-fitting m-dimensional representation is given by the m eigenvectors of $\mathbf{B}$ corresponding to the m largest eigenvalues. The adequacy of the m-dimensional representation can be judged by the size of the criterion

$$P_m = \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{n} \lambda_i}.$$

Values of P_m of the order of 0.8 suggest a reasonable fit.

Classical MD Scaling

Measure of fit is how well the proximities match distances in the geometrical model.

When the observed proximity matrix is not Euclidean, the matrix $\mathbf{B}$ is not positive-definite. In such cases, some of the eigenvalues of $\mathbf{B}$ will be negative; correspondingly, some coordinate values will be complex numbers. If, however, $\mathbf{B}$ has only a small number of small negative eigenvalues, a useful representation of the proximity matrix may still be possible using the eigenvectors associated with the m largest positive eigenvalues. The adequacy of the resulting solution might be assessed using one of the following two criteria suggested by Mardia et al. (1979):

$$P_m^{(1)} = \frac{\sum_{i=1}^{m} |\lambda_i|}{\sum_{i=1}^{n} |\lambda_i|},$$

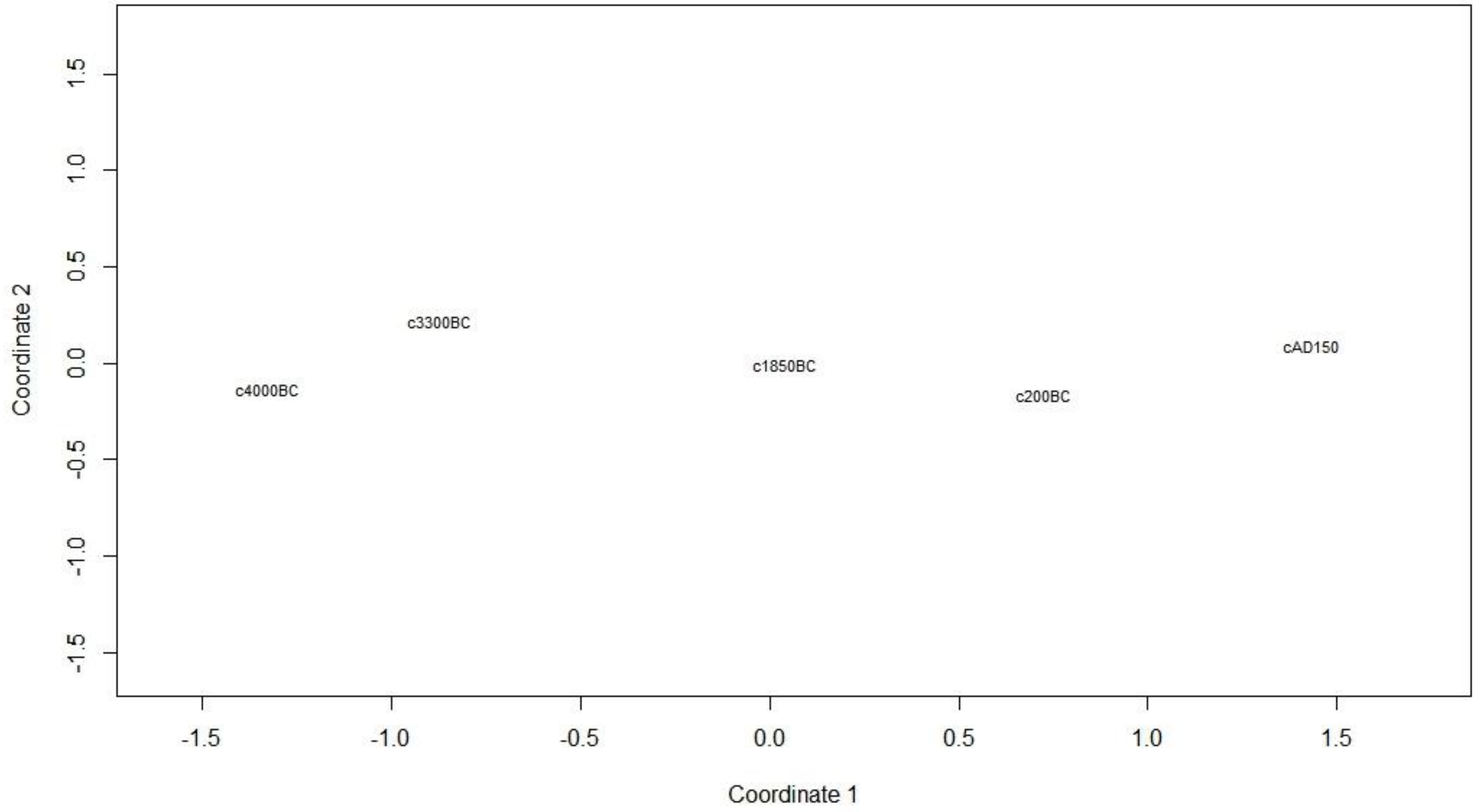
$$P_m^{(2)} = \frac{\sum_{i=1}^{m} \lambda_i^2}{\sum_{i=1}^{n} \lambda_i^2}.$$

Example: `skulls` data

Table 4.2: `skulls` data. Measurements of four variables taken from Egyptian skulls of five periods.

epoch	mb	bh	bl	nh	epoch	mb	bh	bl	nh	epoch	mb	bh	bl	nh
c4000BC	131	138	89	49	c3300BC	137	136	106	49	c200BC	132	133	90	53
c4000BC	125	131	92	48	c3300BC	126	131	100	48	c200BC	134	134	97	54
c4000BC	131	132	99	50	c3300BC	135	136	97	52	c200BC	135	135	99	50
c4000BC	119	132	96	44	c3300BC	129	126	91	50	c200BC	133	136	95	52
c4000BC	136	143	100	54	c3300BC	134	139	101	49	c200BC	136	130	99	55
c4000BC	138	137	89	56	c3300BC	131	134	90	53	c200BC	134	137	93	52
c4000BC	139	130	108	48	c3300BC	132	130	104	50	c200BC	131	141	99	55
c4000BC	125	136	93	48	c3300BC	130	132	93	52	c200BC	129	135	95	47
c4000BC	131	134	102	51	c3300BC	135	132	98	54	c200BC	136	128	93	54
c4000BC	134	134	99	51	c3300BC	130	128	101	51	c200BC	131	125	88	48

Plot Solution: `skulls`

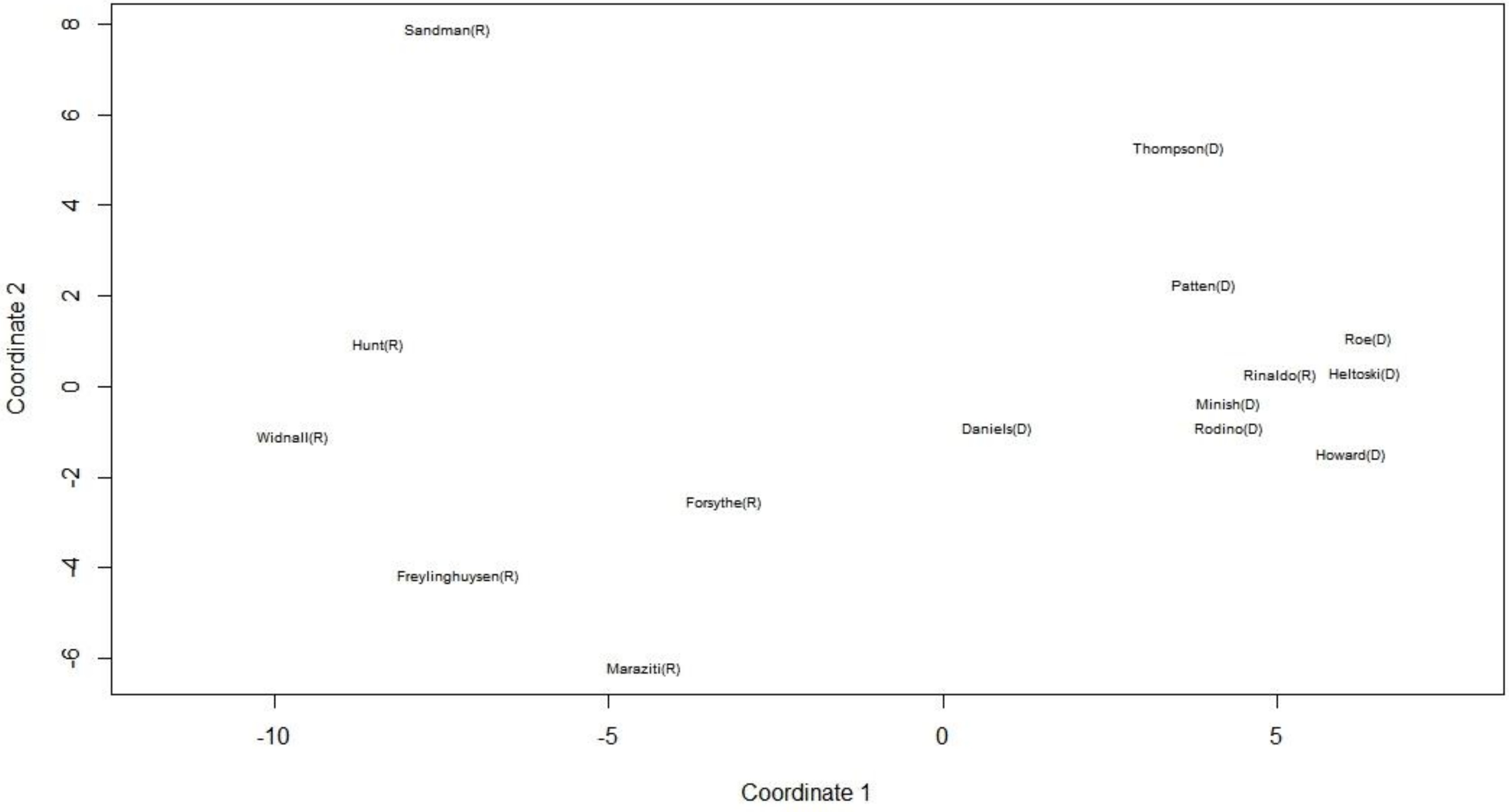


Example: `voting` data

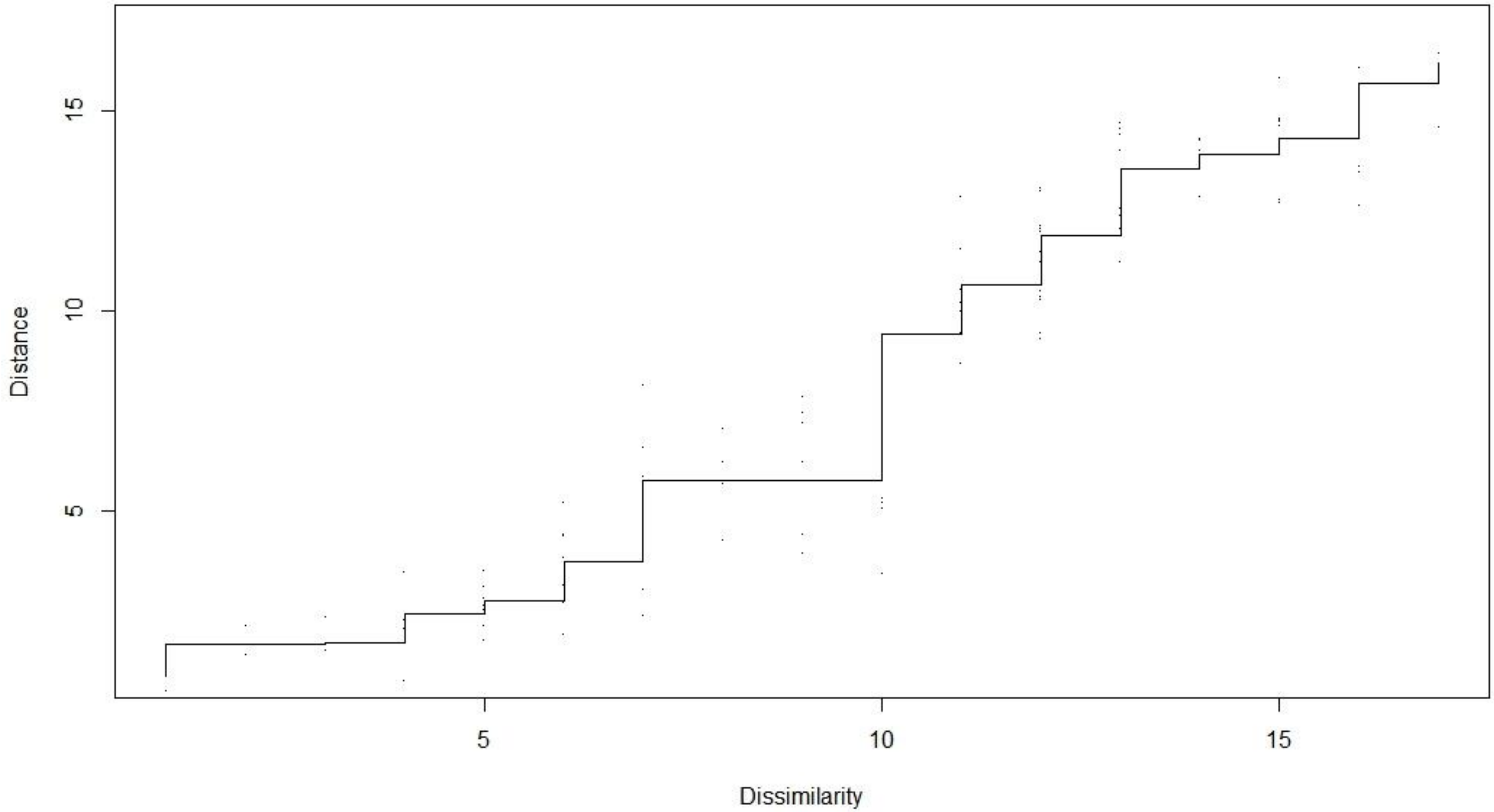
Table 4.4: `voting` data. House of Representatives voting data; (R) is short for Republican, (D) for Democrat.

	Hnt	Snd	Hwr	Thm	Fry	Frs	Wdn	Roe	Hlt	Rdn	Mns	Rnl	Mrz	Dnl	Ptt
Hunt(R)	0														
Sandman(R)	8	0													
Howard(D)	15	17	0												
Thompson(D)	15	12	9	0											
Freylinghuysen(R)	10	13	16	14	0										
Forsythe(R)	9	13	12	12	8	0									
Widnall(R)	7	12	15	13	9	7	0								
Roe(D)	15	16	5	10	13	12	17	0							
Heltoski(D)	16	17	5	8	14	11	16	4	0						
Rodino(D)	14	15	6	8	12	10	15	5	3	0					
Minish(D)	15	16	5	8	12	9	14	5	2	1	0				
Rinaldo(R)	16	17	4	6	12	10	15	3	1	2	1	0			
Maraziti(R)	7	13	11	15	10	6	10	12	13	11	12	12	0		
Daniels(D)	11	12	10	10	11	6	11	7	7	4	5	6	9	0	
Patten(D)	13	16	7	7	11	10	13	6	5	6	5	4	13	9	0

Plot Solution: `voting`



Shepard Plot: `voting`



Example: `WWIILeaders` data

Table 4.5: `WWIIleaders` data. Subjective distances between WWII leaders.

	Htl	Mss	Chr	Esn	Stl	Att	Frn	DGl	MT-	Trm	Chm	Tit
Hitler	0											
Mussolini	3	0										
Churchill	4	6	0									
Eisenhower	7	8	4	0								
Stalin	3	5	6	8	0							
Attlee	8	9	3	9	8	0						
Franco	3	2	5	7	6	7	0					
De Gaulle	4	4	3	5	6	5	4	0				

Plot Solution: `WWIILeaders`