Multidimensional Scaling

Displaying multivariate data in
low-dimensional space

Introduction

- Goal: fit the original data into a low-dimensional coordinate system such that any distortion caused by the reduction in dimensionality is minimized.
- Distortion is measured by the similarities or dissimilarities (distances) between the original data points.
- Such techniques are aka *ordination* of the data.
- Summary: multidimensional scaling is a method to find a representation of N high-dimensional items in a lower dimension such that the new distances “nearly match” the original distances between the items.

Classical MDS

Let $\boldsymbol{X}$ be a $n \times p$ data matrix. Define $\boldsymbol{B} = \boldsymbol{XX'}$. Then the Euclidean distances between rows of $\boldsymbol{X}$ can be written as:

$$d_{ij}^2 = b_{ii} + b_{jj} - 2b_{ij}$$

<u>Idea</u>: Solve for b_{ij} if you know only the d_{ij}

<u>Solution</u>: $b_{ij} = -\frac{1}{2}\left(d_{ij}^2 - d_{i\cdot}^2 - d_{\cdot j}^2 + d_{\cdot\cdot}^2\right)$

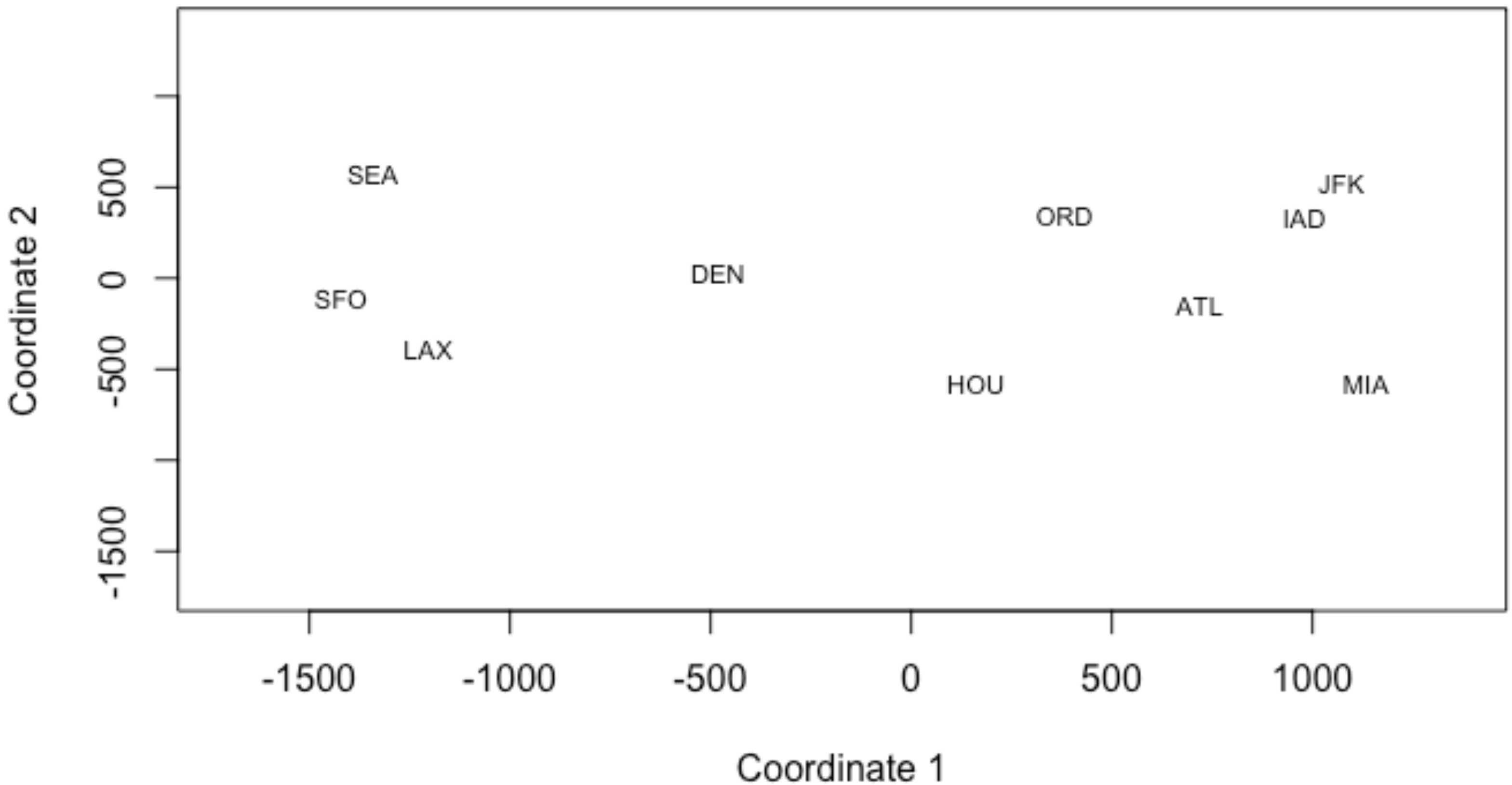
Finally, factor $\boldsymbol{B}$ as $\boldsymbol{B} = \boldsymbol{V\Lambda V'}$ and derive

$$\boldsymbol{X} = \boldsymbol{V\Lambda}^{1/2}$$

Example: Airline Distances

	ATL	ORD	DEN	HOU	LAX	MIA	JFK	SFO	SEA
ORD	587								
DEN	1212	920							
HOU	701	940	879						
LAX	1936	1745	831	1374					
MIA	604	1188	1726	968	2339				
JFK	748	713	1631	1420	2451	1092			
SFO	2139	1858	949	1645	347	2594	2571		
SEA	2182	1737	1021	1891	959	2734	2408	678	
IAD	543	597	1494	1220	2300	923	205	2442	2329

Plot of the MDS results



The Basic Non-Metric Algorithm

For N items, there are

$$M = \frac{N(N-1)}{2}$$

total similarities (distances) between all possible pairs of different items.

Assume no ties and arrange the similarities in ascending order:

$$s_{i_1 j_1} < s_{i_2 j_2} < \cdots < s_{i_M j_M}$$

That is, the pair $i_1 j_1$ is the least similar and $i_M j_M$ is the most similar pair. We want to find a q-dimensional representation of the N items such that the new distances, $d_{ij}^{(q)}$, match the original ordering. A perfect match occurs when:

$$d_{i_1 j_1}^{(q)} > d_{i_2 j_2}^{(q)} > \cdots > d_{i_M j_M}^{(q)}$$

Measuring the Fit

Kruskal proposed a measure of the extent to which a geometrical representation falls short of a perfect match. It is denoted the stress:

$$\text{Stress}(q) = \sqrt{\frac{\sum_{i<j}\left(d_{ij}^{(q)} - \hat{d}_{ij}^{(q)}\right)^2}{\sum_{i<j}\left(d_{ij}^{(q)}\right)^2}}$$

where $\hat{d}_{ij}^{(q)}$ are numbers known to satisfy the monotonicity property.

Idea: Find a representation such that the stress is as small as possible.

Guidelines:

Stress	Goodness of fit
20%	Poor
10%	Fair
5%	Good
2.5%	Excellent
0%	Perfect

Algorithm Steps

1. Obtain the $N(N\text{-}1)/2$ similarities and order them.
2. Using q dimensions determine initial distances $d_{ij}^{(q)}$. Choose numbers $\hat{d}_{ij}^{(q)}$ such that they are monotone and minimize the stress statistic. The software usually uses some sort of monotone regression method to produce fitted distances.
3. Using the $\hat{d}_{ij}^{(q)}$ from the previous step find a new projection with improved $d_{ij}^{(q)}$ that minimize the stress further. Repeat steps 1 and 2 until convergence.
4. Plot stress(q) vs. q and choose the best dimension q.

Example: Voting

	Hunt(R)	Sandman(R)	Howard(D)	Thompson(D)	Freylinghuysen(R)	Forsythe(R)	Widnall(R)
Hunt(R)	0	8	15	15	10	9	7
Sandman(R)	8	0	17	12	13	13	12
Howard(D)	15	17	0	9	16	12	15
Thompson(D)	15	12	9	0	14	12	13
Freylinghuysen(R)	10	13	16	14	0	8	9
Forsythe(R)	9	13	12	12	8	0	7
Widnall(R)	7	12	15	13	9	7	0
Roe(D)	15	16	5	10	13	12	17
Heltoski(D)	16	17	5	8	14	11	16
Rodino(D)	14	15	6	8	12	10	15
Minish(D)	15	16	5	8	12	9	14
Rinaldo(R)	16	17	4	6	12	10	15
Maraziti(R)	7	13	11	15	10	6	10
Daniels(D)	11	12	10	10	11	6	11
Patten(D)	13	16	7	7	11	10	13

	Roe(D)	Heltoski(D)	Rodino(D)	Minish(D)	Rinaldo(R)	Maraziti(R)	Daniels(D)	Patten(D)
Hunt(R)	15	16	14	15	16	7	11	13
Sandman(R)	16	17	15	16	17	13	12	16
Howard(D)	5	5	6	5	4	11	10	7
Thompson(D)	10	8	8	8	6	15	10	7
Freylinghuysen(R)	13	14	12	12	12	10	11	11
Forsythe(R)	12	11	10	9	10	6	6	10
Widnall(R)	17	16	15	14	15	10	11	13
Roe(D)	0	4	5	5	3	12	7	6
Heltoski(D)	4	0	3	2	1	13	7	5
Rodino(D)	5	3	0	1	2	11	4	6
Minish(D)	5	2	1	0	1	12	5	5
Rinaldo(R)	3	1	2	1	0	12	6	4
Maraziti(R)	12	13	11	12	12	0	9	13
Daniels(D)	7	7	4	5	6	9	0	9
Patten(D)	6	5	6	5	4	13	9	0

Plot of results