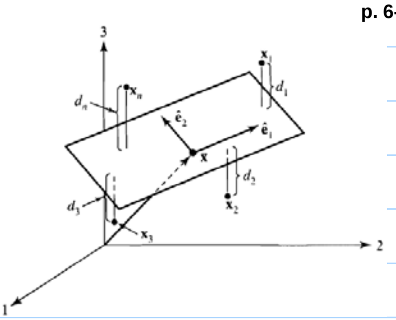
Multidimensional Scaling (MDS)



- Recall: Principal component scores
- observed data in MDS: proximity matrix

$$\underset{(n\times p)}{\mathbf{X}} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

PCA ↓

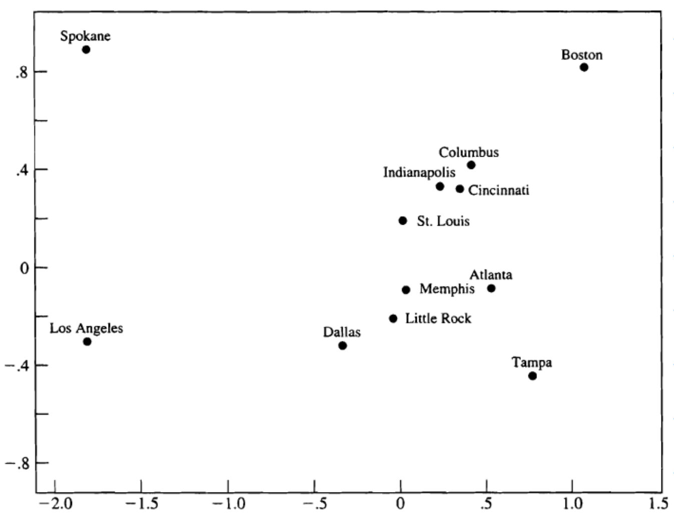
$$Y = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1q} \\ y_{21} & y_{22} & \cdots & y_{2q} \\ \cdots & \cdots & & \cdots \\ y_{n1} & y_{n2} & \cdots & y_{nq} \end{bmatrix}$$

A proximity matrix $\mathbf{D}=[d_{ij}]$ represents the "distance" (similarity/dissimilarity) between rows or columns of $\mathbf{X}$

- Q: Suppose that $\mathbf{X}$ is not observed. Given $\mathbf{D}$, how do we find $\mathbf{X}$ (or $\mathbf{Y}$)?
- Note: a proximity matrix is invariant to (1) change in location, (2) rotation, (3) reflections $\Rightarrow$ cannot expect to recover $\mathbf{X}$ completely
- Some suggestions to convert similarity matrix $\mathbf{S}=[s_{ij}]$ to dissimilarity matrix $\mathbf{D}$
  - d_{ij} = constant – s_{ij}
  - $d_{ij} = 1/s_{ij}$ – constant
  - $d_{ij} = s_{ii} + s_{jj} - 2\, s_{ij}$

- Example of proximity matrix:
  - airline distance data

	Atlanta (1)	Boston (2)	Cincinnati (3)	Columbus (4)	Dallas (5)	Indianapolis (6)	Little Rock (7)	Los Angeles (8)	Memphis (9)	St. Louis (10)	Spokane (11)	Tampa (12)
(1)	0											
(2)	1068	0										
(3)	461	867	0									
(4)	549	769	107	0								
(5)	805	1819	943	1050	0							
(6)	508	941	108	172	882	0						
(7)	505	1494	618	725	325	562	0					
(8)	2197	3052	2186	2245	1403	2080	1701	0				
(9)	366	1355	502	586	464	436	137	1831	0			
(10)	558	1178	338	409	645	234	353	1848	294	0		
(11)	2467	2747	2067	2131	1891	1959	1988	1227	2042	1820	0	
(12)	467	1379	928	985	1077	975	912	2480	779	1016	2821	0



- correlation matrix of crime dataset

	Murder	Rape	Robbery	Assault	Burglary	Larceny	MVT
Murder	1.0000000	0.2260129	0.6991807	0.5022187	0.5129268	0.1642008	0.4784107
Rape	0.2260129	1.0000000	0.2390043	0.2937968	0.5179029	0.2972498	0.1771553
Robbery	0.6991807	0.2390043	1.0000000	0.6605061	0.5961077	0.1519019	0.7056040
Assault	0.5022187	0.2937968	0.6605061	1.0000000	0.6818327	0.5113177	0.5753717
Burglary	0.5129268	0.5179029	0.5961077	0.6818327	1.0000000	0.5891757	0.5771577
Larceny	0.1642008	0.2972498	0.1519019	0.5113177	0.5891757	1.0000000	0.2166964
MVT	0.4784107	0.1771553	0.7056040	0.5753717	0.5771577	0.2166964	1.0000000

- survey of legal offences (1.Assault and battery, 2.Rape, 3.Embezzlement, 4.Perjury, 5.Libel, 6.Burglary, 7.Prostitution, 8.Receiving stolen goods)

0							
21.1	0						
71.2	54.1	0					
36.4	36.4	36.4	0				
52.1	54.1	52.1	0.7	0			
89.9	75.2	36.4	54.1	53.0	0		
53.0	73.0	75.2	52.1	36.4	88.3	0	
90.1	93.2	71.2	63.4	52.1	36.4	73.0	0

➢ A proximity matrix is called

- *distance-like* if (1) $d_{ij} \geq 0$, (2) $d_{ii} = 0$, (3) $d_{ij} = d_{ji}$
- *metric* if in addition to (1)(2)(3), it satisfies $d_{ij} \leq d_{ik} + d_{jk}$
- *Euclidean* if there exists a configuration of points in Euclidean space with distance(point$_i$, point$_j$)=d_{ij}

- classical multidimensional scaling

➢ assume that the *observed* $n \times n$ proximity matrix **D** is a matrix of Euclidean distances derived from a raw $n \times q$ data matrix, **X**, which is *not observed*.

➢ define an $n \times n$ matrix **B**

$$\mathbf{B} = \mathbf{X}\mathbf{X}' (= (\mathbf{X}\mathbf{M})(\mathbf{X}\mathbf{M})')$$

M: an orthogonal matrix

- the elements of **B** are given by

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

➢ the squared Euclidean distances between the rows of **X** can be written in terms of the elements of **B** as

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

- idea: If the b_{ij}'s could be found in terms of the d_{ij}'s in the equation above, then we can derive **X** from **B** by factoring **B**.
- to obtain **B** from **D**, no unique solution exists unless a location constraint is introduced. Usually, the center of the columns of **X** are set at origin, i.e.,

$$\sum_{i=1}^{n} x_{ik} = 0, \quad \text{for all } k$$

- these constraints imply that sum of the terms in any row of **B** must be 0, i.e.,

$$\sum_{j=1}^{n} b_{ij} = \sum_{j=1}^{n} \sum_{k=1}^{q} x_{ik} x_{jk} = \sum_{k=1}^{q} x_{ik} \left(\sum_{j=1}^{n} x_{jk} \right)$$

- Let T be the trace of **B**. To obtain **B** from **D**, notice that
  - $\sum_{i=1}^{n} d_{ij}^2 = T + nb_{jj}$
  - $\sum_{j=1} d_{ij}^2 = nb_{ii} + T$
  - $\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2 = 2nT$
- the elements of **B** can be found from **D** as

$$b_{ij} = -\frac{1}{2}\left[d_{ij}^2 - d_{i\cdot}^2 - d_{\cdot j}^2 + d_{\cdot\cdot}^2\right]$$

where $d_{i\cdot}^2 = (\sum_{j=1}^{n} d_{ij}^2)/n$, $d_{\cdot j}^2 = (\sum_{i=1}^{n} d_{ij}^2)/n$, $d_{\cdot\cdot}^2 = (\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2)/n^2$

- $\mathbf{B}$ can be written as

$$\mathbf{B} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}',$$

where $\mathbf{\Lambda} = diag[\lambda_1, \cdots, \lambda_n]$ $(\lambda_1 \geq \cdots \geq \lambda_n)$ is the diagonal matrix of eigenvalues of $\mathbf{B}$ and $\mathbf{V} = [\mathbf{V_1}, \cdots, \mathbf{V_n}]$ is the corresponding matrix of normalized eigenvectors (i.e., $\mathbf{V_i'}\mathbf{V_i} = 1$)

  - Note: when **D** arises from an n×q data matrix, the rank of **B** is q (i.e, the last n–q eigenvalues should be zero)
  - So, $\mathbf{B}$ can be chosen as

$$\mathbf{B} = \mathbf{V}^*\mathbf{\Lambda}^*\mathbf{V}^{*\prime},$$

where $\mathbf{V}^*$ contains the first q eigenvectors and $\mathbf{\Lambda}^*$ the first q eigenvalues

  - Thus, a solution of **X** is $\mathbf{X} = \mathbf{V}^*\mathbf{\Lambda}^{*1/2}$

- The adequacy of the q-dimensional representation can be judged by the size of the criterion

$$(\textstyle\sum_{i=1}^{q} \lambda_i)/(\sum_{i=1}^{n} \lambda_i)$$

- When the observed proximity matrix is not Euclidean, the matrix **B** is not positive-definite. In such case, some of the eigenvalues of **B** will be negative; correspondingly, some coordinate values will be complex numbers.
  - If **B** has only a small number of small negative eigenvalues, it's still possible to use the eigenvectors associated with the q largest positive eigenvalues
  - adequacy of the resulting solution might be assessed using
    - $(\sum_{i=1}^{q} |\lambda_i|)/(\sum_{i=1}^{n} |\lambda_i|)$
    - $(\sum_{i=1}^{q} \lambda_i^2)/(\sum_{i=1}^{n} \lambda_i^2)$

- Some other issues
  - metric scaling (define loss function for **D** and the distance matrix based on **X**, called *stress*, and find **X** to minimize stress)
  - non-metric scaling (applied when the actual values of **D** is not reliable, but their orders can be trusted)
  - 3-way multidimensional scaling (proximity matrix result from individual assessments of dissimilarity and more than one individual is sampled)
  - asymmetric proximity matrix ($d_{ij} \neq d_{ji}$)

❖ **Reading**: Reference, 5.2