Multidimensional Scaling (MDS)

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

\[ X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \]

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

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

\[ d_{ij} = \text{constant} - s_{ij} \quad d_{ij} = 1/s_{ij} - \text{constant} \]
\[ d_{ij} = s_{ii} + s_{jj} - 2 s_{ij} \]

- Note: a proximity matrix is invariant to (1) change in location, (2) rotation, (3) reflections \( \Rightarrow \) cannot expect to recover \( X \) completely

Example of proximity matrix:

<table>
<thead>
<tr>
<th></th>
<th>Atlanta</th>
<th>Boston</th>
<th>Cincinnati</th>
<th>Columbus</th>
<th>Dallas</th>
<th>Indianapolis</th>
<th>Little Rock</th>
<th>Los Angeles</th>
<th>Memphis</th>
<th>St. Louis</th>
<th>Spokane</th>
<th>Tampa</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>0</td>
<td>1048</td>
<td>0</td>
<td>461</td>
<td>549</td>
<td>805</td>
<td>508</td>
<td>565</td>
<td>2197</td>
<td>366</td>
<td>558</td>
<td>2467</td>
</tr>
<tr>
<td>(2)</td>
<td>1048</td>
<td>0</td>
<td>943</td>
<td>108</td>
<td>769</td>
<td>1819</td>
<td>941</td>
<td>1494</td>
<td>2186</td>
<td>1355</td>
<td>1178</td>
<td>2747</td>
</tr>
<tr>
<td>(3)</td>
<td>461</td>
<td>943</td>
<td>0</td>
<td>108</td>
<td>618</td>
<td>905</td>
<td>108</td>
<td>618</td>
<td>1245</td>
<td>502</td>
<td>338</td>
<td>2237</td>
</tr>
<tr>
<td>(4)</td>
<td>549</td>
<td>108</td>
<td>108</td>
<td>172</td>
<td>725</td>
<td>941</td>
<td>108</td>
<td>725</td>
<td>1403</td>
<td>586</td>
<td>409</td>
<td>645</td>
</tr>
<tr>
<td>(5)</td>
<td>805</td>
<td>108</td>
<td>108</td>
<td>882</td>
<td>325</td>
<td>1819</td>
<td>941</td>
<td>1494</td>
<td>2186</td>
<td>464</td>
<td>409</td>
<td>645</td>
</tr>
<tr>
<td>(6)</td>
<td>508</td>
<td>108</td>
<td>108</td>
<td>882</td>
<td>325</td>
<td>941</td>
<td>108</td>
<td>725</td>
<td>1403</td>
<td>464</td>
<td>409</td>
<td>645</td>
</tr>
<tr>
<td>(7)</td>
<td>565</td>
<td>1494</td>
<td>1494</td>
<td>562</td>
<td>0</td>
<td>2197</td>
<td>366</td>
<td>558</td>
<td>2467</td>
<td>137</td>
<td>353</td>
<td>1848</td>
</tr>
<tr>
<td>(8)</td>
<td>2197</td>
<td>2186</td>
<td>2186</td>
<td>0</td>
<td>0</td>
<td>1701</td>
<td>1831</td>
<td>1848</td>
<td>294</td>
<td>1701</td>
<td>1831</td>
<td>0</td>
</tr>
<tr>
<td>(9)</td>
<td>366</td>
<td>1355</td>
<td>502</td>
<td>464</td>
<td>464</td>
<td>1701</td>
<td>1831</td>
<td>1848</td>
<td>0</td>
<td>1701</td>
<td>1831</td>
<td>0</td>
</tr>
<tr>
<td>(10)</td>
<td>558</td>
<td>1178</td>
<td>338</td>
<td>409</td>
<td>409</td>
<td>1701</td>
<td>1831</td>
<td>1848</td>
<td>594</td>
<td>1701</td>
<td>1831</td>
<td>0</td>
</tr>
<tr>
<td>(11)</td>
<td>2467</td>
<td>2747</td>
<td>2237</td>
<td>294</td>
<td>294</td>
<td>1701</td>
<td>1831</td>
<td>1848</td>
<td>294</td>
<td>1701</td>
<td>1831</td>
<td>0</td>
</tr>
<tr>
<td>(12)</td>
<td>467</td>
<td>1379</td>
<td>586</td>
<td>464</td>
<td>464</td>
<td>1701</td>
<td>1831</td>
<td>1848</td>
<td>594</td>
<td>1701</td>
<td>1831</td>
<td>0</td>
</tr>
</tbody>
</table>

Example of proximity matrix:

- airline distance data

- Note: a proximity matrix is invariant to (1) change in location, (2) rotation, (3) reflections \( \Rightarrow \) cannot expect to recover \( X \) completely

Some suggestions to convert similarity matrix \( S = [s_{ij}] \) to dissimilarity matrix \( D \)

\[ d_{ij} = \text{constant} - s_{ij} \quad d_{ij} = 1/s_{ij} - \text{constant} \]
\[ d_{ij} = s_{ii} + s_{jj} - 2 s_{ij} \]
correlation matrix of crime dataset

<table>
<thead>
<tr>
<th></th>
<th>Murder</th>
<th>Rape</th>
<th>Robbery</th>
<th>Assault</th>
<th>Burglary</th>
<th>Larceny</th>
<th>MVT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Murder</td>
<td>1.000000</td>
<td>0.2260129</td>
<td>0.6991807</td>
<td>0.5022187</td>
<td>0.5129268</td>
<td>0.1642008</td>
<td>0.4784107</td>
</tr>
<tr>
<td>Rape</td>
<td>0.2260129</td>
<td>1.000000</td>
<td>0.2390043</td>
<td>0.2937968</td>
<td>0.5179029</td>
<td>0.2972498</td>
<td>0.1771553</td>
</tr>
<tr>
<td>Robbery</td>
<td>0.6991807</td>
<td>0.2390043</td>
<td>1.000000</td>
<td>0.6605061</td>
<td>0.5961077</td>
<td>0.1519019</td>
<td>0.7056040</td>
</tr>
<tr>
<td>Assault</td>
<td>0.5022187</td>
<td>0.2937968</td>
<td>0.6605061</td>
<td>1.000000</td>
<td>0.6818327</td>
<td>0.5133177</td>
<td>0.5753717</td>
</tr>
<tr>
<td>Burglary</td>
<td>0.5129268</td>
<td>0.5179029</td>
<td>0.5961077</td>
<td>0.6818327</td>
<td>1.000000</td>
<td>0.5891757</td>
<td>0.5771577</td>
</tr>
<tr>
<td>Larceny</td>
<td>0.1642008</td>
<td>0.2972498</td>
<td>0.1519019</td>
<td>0.5133177</td>
<td>0.5891757</td>
<td>1.000000</td>
<td>0.2166964</td>
</tr>
<tr>
<td>MVT</td>
<td>0.4784107</td>
<td>0.1771553</td>
<td>0.7056040</td>
<td>0.5753717</td>
<td>0.5771577</td>
<td>0.2166964</td>
<td>1.000000</td>
</tr>
</tbody>
</table>

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 \( B \)
    \[ B = XX' = (XM)(XM)' \]
    \( M \): 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

- $\sum_{i=1}^{n} d_{ij}^2 = T + nb_{jj}$
- $\sum_{j=1}^{n} d_{ij}^2 = nb_{ii} + T$
- $\sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}^2 = 2nT$

The elements of $B$ can be found from $D$ as

$$b_{ij} = -\frac{1}{2} [d_{ij}^2 - d_{ii}^2 - d_{jj}^2 + d_{ij}^2]$$

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

$B$ can be written as

$$B = V\Lambda V',$$

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

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

$$B = V^*\Lambda^*V'^*,$$

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

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

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

$$\frac{\sum_{i=1}^{q} \lambda_i}{\sum_{i=1}^{n} \lambda_i}$$

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

  - $\frac{\sum_{i=1}^{q} |\lambda_i|}{\sum_{i=1}^{n} |\lambda_i|}$
  - $\frac{\sum_{i=1}^{q} \lambda_i^2}{\sum_{i=1}^{n} \lambda_i^2}$

- 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 these 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