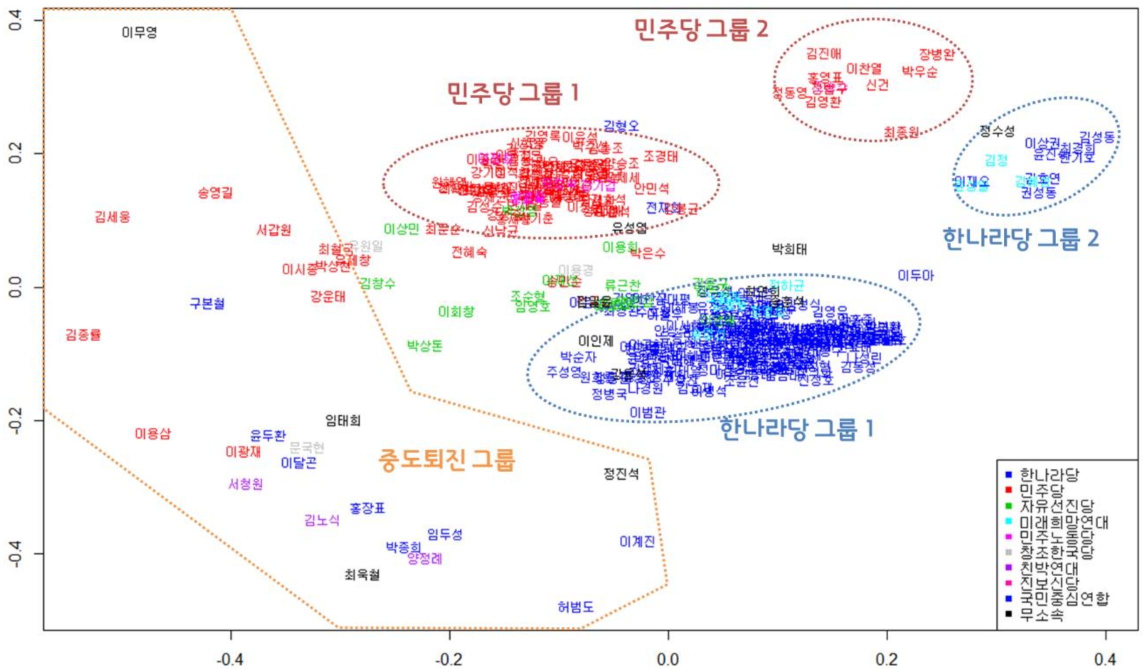
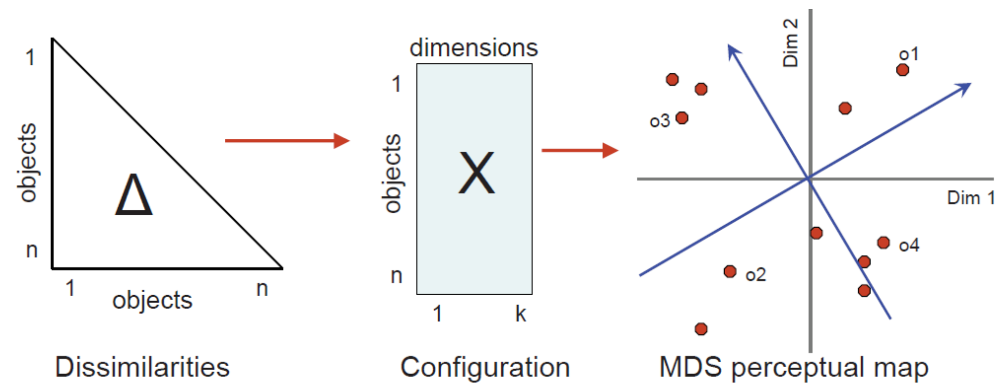


Dimensionality Reduction: Multidimensional Scaling

Pilsung Kang
School of Industrial Management Engineering
Korea University

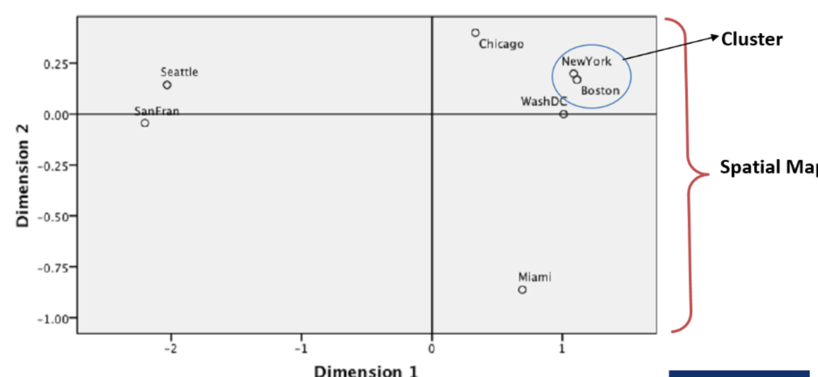
Multidimensional Scaling: MDS

- Multidimensional Scaling
  - ✓ Aims to place each object in D-dimensional space such that the between-object distances are preserved as well as possible



Distances Matrix: Symmetric

		1 BOST	2 NY	3 DC	4 MIAM	5 CHIC	6 SEAT	7 SF	8 LA	9 DENV
1	BOSTON	0	206	429	1504	963	2976	3095	2979	1949
2	NY	206	0	233	1308	802	2815	2934	2786	1771
3	DC	429	233	0	1075	671	2684	2799	2631	1616
4	MIAMI	1504	1308	1075	0	1329	3273	3053	2687	2037
5	CHICAGO	963	802	671	1329	0	2013	2142	2054	996
6	SEATTLE	2976	2815	2684	3273	2013	0	808	1131	1307
7	SF	3095	2934	2799	3053	2142	808	0	379	1235
8	LA	2979	2786	2631	2687	2054	1131	379	0	1059
9	DENVER	1949	1771	1616	2037	996	1307	1235	1059	0



Multidimensional Scaling: MDS

- PCA vs. MDS

	Principal Component Analysis (PCA)	Multidimensional Scaling (MDS)
Data	n objects in a d-dimensional space ($\mathbf{X}$ in R^d)	Proximity matrix between n objects (n by n matrix $\mathbf{D}$)
Purpose	Find a set of bases to preserve the original variance	Find a set of coordinates that preserve the distance information between objects
Output	1. d bases (eigenvectors, PCs) 2. d eigenvalues	Coordinate of each object in d-dimension ($\mathbf{X}$ in R^d)

MDS Procedure

- Step 1: Construct Proximity/Distance Matrix
  - ✓ If there exist the coordinates for the objects, compute the similarity/distance between them

$$distance-like \;\; \text{if } (1)\; d_{ij} \geq 0, \quad (2)\; d_{ii} = 0, \quad (3)\; d_{ij} = d_{ji}$$

$$metric \;\; \text{if in addition to } (1),\; (2),\; (3),\; \text{it satisfies } d_{ij} \leq d_{ik} + d_{jk}$$

  - ✓ Distance: Euclidean, Manhattan, etc.
  - ✓ Similarity: Correlation, Jaccard, etc.

X (d by n)

	x_1	x_2	x_3	x_4	…	x_n
v_1						
v_2						
v_3						
…						
v_d						

D (n by n)

	x_1	x_2	x_3	x_4	…	x_n
x_1						
x_2						
x_3						
x_4						
…						
x_n						

$$d_{rs}^2 = (\mathbf{x}_r - \mathbf{x}_s)^T (\mathbf{x}_r - \mathbf{x}_s)$$

MDS Procedure

- Step 2: Extract the coordinates that preserve the distance information
  - ✓ Each element of the distance matrix **D** can be expressed as

$$d_{rs}^2 = (\mathbf{x}_r - \mathbf{x}_s)^T(\mathbf{x}_r - \mathbf{x}_s)$$

  - ✓ Inner product matrix **B** can be obtained from the distance matrix **D**

$$[\mathbf{B}]_{rs} = b_{rs} = \mathbf{x}_r^T \mathbf{x}_s$$

    - Assume that the means of all p variables are 0

$$\sum_{r=1}^{n} x_{ri} = 0 \ , \ (i = 1, 2, ..., p)$$

$$d_{rs}^2 = \mathbf{x}_r^T \mathbf{x}_r + \mathbf{x}_s^T \mathbf{x}_s - 2\mathbf{x}_r^T \mathbf{x}_s$$

MDS Procedure

- Step 2: Extract the coordinates that preserve the distance information

$$d_{rs}^2 = (\mathbf{x}_r - \mathbf{x}_s)^T(\mathbf{x}_r - \mathbf{x}_s)$$

$$\frac{1}{n}\sum_{r=1}^{n} d_{rs}^2 = \frac{1}{n}\sum_{r=1}^{n} \mathbf{x}_r^T\mathbf{x}_r + \frac{1}{n}\sum_{r=1}^{n} \mathbf{x}_s^T\mathbf{x}_s - \frac{2}{n}\sum_{r=1}^{n} \mathbf{x}_r^T\mathbf{x}_s = \frac{1}{n}\sum_{r=1}^{n} \mathbf{x}_r^T\mathbf{x}_r + \mathbf{x}_s^T\mathbf{x}_s$$

$$\mathbf{x}_s^T\mathbf{x}_s = \frac{1}{n}\sum_{r=1}^{n} d_{rs}^2 - \frac{1}{n}\sum_{r=1}^{n} \mathbf{x}_r^T\mathbf{x}_r$$

$$\frac{1}{n}\sum_{s=1}^{n} d_{rs}^2 = \frac{1}{n}\sum_{s=1}^{n} \mathbf{x}_r^T\mathbf{x}_r + \frac{1}{n}\sum_{s=1}^{n} \mathbf{x}_s^T\mathbf{x}_s - \frac{2}{n}\sum_{s=1}^{n} \mathbf{x}_r^T\mathbf{x}_s = \mathbf{x}_r^T\mathbf{x}_r + \frac{1}{n}\sum_{s=1}^{n} \mathbf{x}_s^T\mathbf{x}_s$$

$$\mathbf{x}_r^T\mathbf{x}_r = \frac{1}{n}\sum_{s=1}^{n} d_{rs}^2 - \frac{1}{n}\sum_{s=1}^{n} \mathbf{x}_s^T\mathbf{x}_s$$

MDS Procedure

- Step 2: Extract the coordinates that preserve the distance information

$$d_{rs}^2 = (\mathbf{x}_r - \mathbf{x}_s)^T (\mathbf{x}_r - \mathbf{x}_s)$$

$$\frac{1}{n^2} \sum_{r=1}^{n} \sum_{s=1}^{n} d_{rs}^2 = \frac{1}{n^2} \sum_{r=1}^{n} \sum_{s=1}^{n} \mathbf{x}_r^T \mathbf{x}_r + \frac{1}{n^2} \sum_{r=1}^{n} \sum_{s=1}^{n} \mathbf{x}_s^T \mathbf{x}_s - \frac{2}{n^2} \sum_{r=1}^{n} \sum_{s=1}^{n} \mathbf{x}_r^T \mathbf{x}_s$$

$$= \frac{1}{n} \sum_{r=1}^{n} \mathbf{x}_r^T \mathbf{x}_r + \frac{1}{n} \sum_{s=1}^{n} \mathbf{x}_s^T \mathbf{x}_s = \frac{2}{n} \sum_{r=1}^{n} \mathbf{x}_r^T \mathbf{x}_r$$

$$\frac{2}{n} \sum_{r=1}^{n} \mathbf{x}_r^T \mathbf{x}_r = \frac{1}{n^2} \sum_{r=1}^{n} \sum_{s=1}^{n} d_{rs}^2$$

MDS Procedure

- Step 2: Extract the coordinates that preserve the distance information

$$b_{rs} = \mathbf{x}_r^T \mathbf{x}_s \qquad (d_{rs}^2 = \mathbf{x}_r^T \mathbf{x}_r + \mathbf{x}_s^T \mathbf{x}_s - 2\mathbf{x}_r^T \mathbf{x}_s)$$

$$= -\frac{1}{2}(d_{rs}^2 - \mathbf{x}_r^T \mathbf{x}_r - \mathbf{x}_s^T \mathbf{x}_s)$$

$$\boxed{\mathbf{x}_s^T \mathbf{x}_s = \frac{1}{n}\sum_{r=1}^{n} d_{rs}^2 - \frac{1}{n}\sum_{r=1}^{n} \mathbf{x}_r^T \mathbf{x}_r \quad \mathbf{x}_r^T \mathbf{x}_r = \frac{1}{n}\sum_{s=1}^{n} d_{rs}^2 - \frac{1}{n}\sum_{s=1}^{n} \mathbf{x}_s^T \mathbf{x}_s}$$

$$= -\frac{1}{2}\left(d_{rs}^2 - \frac{1}{n}\sum_{s=1}^{n} d_{rs}^2 + \frac{1}{n}\sum_{s=1}^{n} \mathbf{x}_s^T \mathbf{x}_s - \frac{1}{n}\sum_{r=1}^{n} d_{rs}^2 + \frac{1}{n}\sum_{r=1}^{n} \mathbf{x}_r^T \mathbf{x}_r\right)$$

$$= -\frac{1}{2}\left(d_{rs}^2 - \frac{1}{n}\sum_{s=1}^{n} d_{rs}^2 - \frac{1}{n}\sum_{r=1}^{n} d_{rs}^2 + \frac{1}{n^2}\sum_{r=1}^{n}\sum_{s=1}^{n} d_{rs}^2\right)$$

$$\boxed{\frac{2}{n}\sum_{r=1}^{n} \mathbf{x}_r^T \mathbf{x}_r = \frac{1}{n^2}\sum_{r=1}^{n}\sum_{s=1}^{n} d_{rs}^2}$$

$$= a_{rs} - a_{r\cdot} - a_{\cdot s} + a_{\cdot\cdot}$$

$$\text{where } a_{rs} = -\frac{1}{2}d_{rs}^2, \ a_{r\cdot} = \frac{1}{n}\sum_{s} a_{rs}, \ a_{\cdot s} = \frac{1}{n}\sum_{r} a_{rs}, \ a_{\cdot\cdot} = \frac{1}{n^2}\sum_{r}\sum_{s} a_{rs}$$

MDS Procedure

- Step 2: Extract the coordinates that preserve the distance information

$$b_{rs} = \mathbf{x}_r^T \mathbf{x}_s = -\frac{1}{2}\Big(d_{rs}^2 - \frac{1}{n}\sum_{s=1}^{n} d_{rs}^2 - \frac{1}{n}\sum_{r=1}^{n} d_{rs}^2 + \frac{1}{n^2}\sum_{r=1}^{n}\sum_{s=1}^{n} d_{rs}^2\Big)$$

$$= a_{rs} - a_{r\cdot} - a_{\cdot s} + a_{\cdot\cdot}$$

$$\Big(\text{where } a_{rs} = -\frac{1}{2}d_{rs}^2,\ a_{r\cdot} = \frac{1}{n}\sum_s a_{rs},\ a_{\cdot s} = \frac{1}{n}\sum_r a_{rs},\ a_{\cdot\cdot} = \frac{1}{n^2}\sum_r\sum_s a_{rs}\Big)$$

$$[\mathbf{A}]_{rs} = a_{rs} \qquad \mathbf{B} = \mathbf{HAH} \qquad \mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$$

고려대학교 KOREA UNIVERSITY

DSBA Data Science & Business Analytics

MDS Procedure

- Step 2: Extract the coordinates that preserve the distance information
  - ✓ Obtain coordinates of **X** from **B** (**X**: n by p, p < n)

$$\mathbf{B} = \mathbf{X}\mathbf{X}^T \quad rank(\mathbf{B}) = rank(\mathbf{X}\mathbf{X}^T) = rank(\mathbf{X}) = p$$

  - ✓ **B** is symmetric, positive semi-definite and of rank p, so it has p non-negative eigenvalues and (n-p) zero eigenvalues (by Eigen-decomposition)

$$\mathbf{B} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T, \quad \mathbf{\Lambda} = diag(\lambda_1, \lambda_2, ..., \lambda_n), \mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_n]$$

  - ✓ Because of (n-p) zero eigenvalues, **B** can be rewritten as

$$\mathbf{B}_1 = \mathbf{V}_1\mathbf{\Lambda}_1\mathbf{V}_1^T, \quad \mathbf{\Lambda}_1 = diag(\lambda_1, \lambda_2, ..., \lambda_p), \mathbf{V}_1 = [\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_p]$$

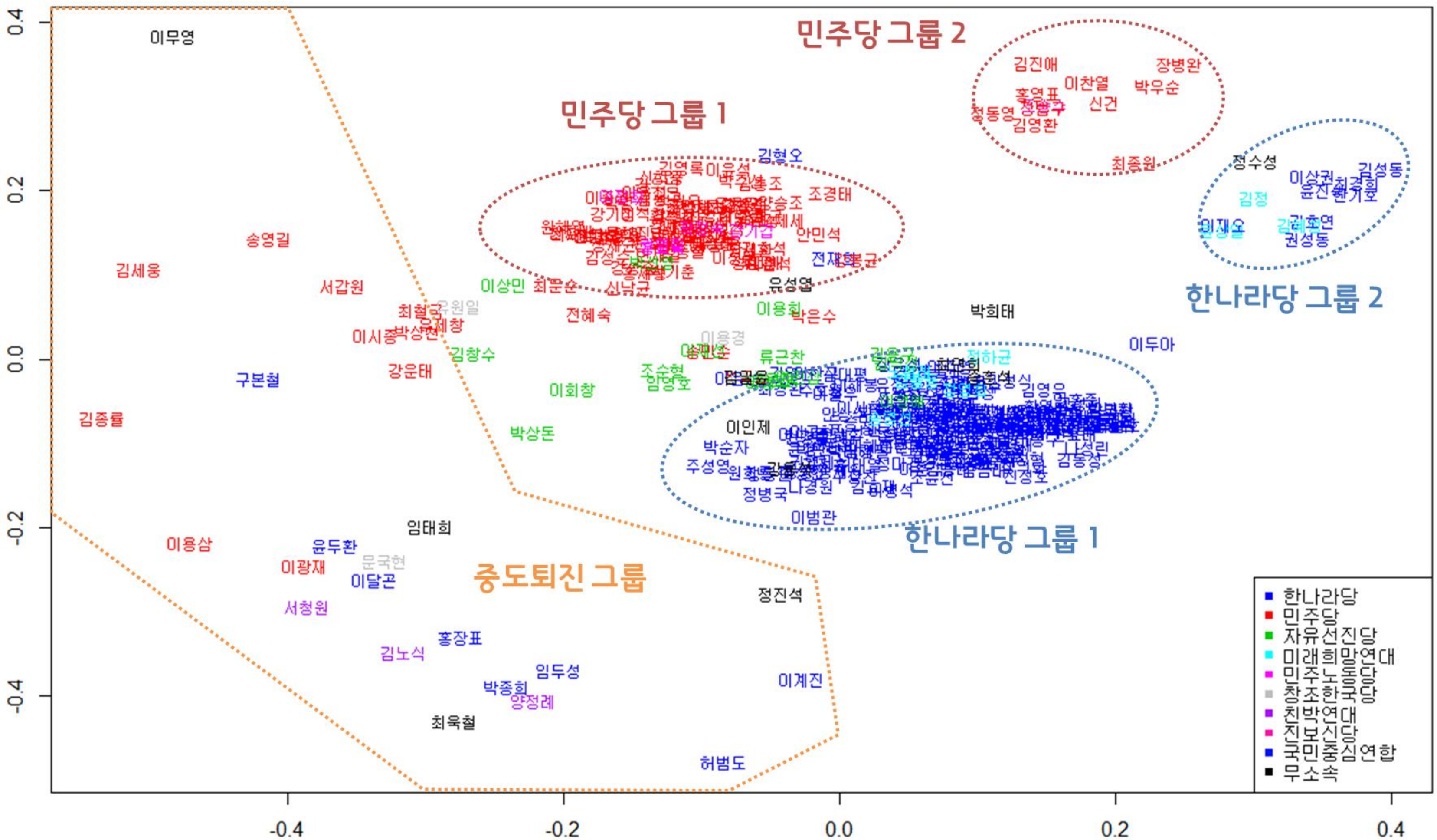
  - ✓ The coordinate matrix **X** is given by

$$\mathbf{X} = \mathbf{V}_1\mathbf{\Lambda}_1^{\frac{1}{2}}$$

고려대학교 KOREA UNIVERSITY

DSBA Data Science & Business Analytics

MDS Example



강필성, 박영준, 조수곤, 김성범. (2013). 대한민국 18대 국회의원 의정활동 분석, 한국경영과학회 추계학술대회


Q & A