Multidimensional Scaling (MDS)

Ying-Chao Hung
Department of Statistics
National Chengchi University
Email: hungy@nccu.edu.tw

Introduction

- Multidimensional scaling (MDS) is a method that represents measures of similarity or dissimilarity among pairs of objects as distances between points in a low-dimensional space (Borg & Groenen, 1997).

- Unsupervised Methods (designed for visualization)
  - Projection Methods: PCA, projection pursuit, etc.
  - MDS
  - Cluster analysis

Define Dissimilarity

- Data matrix:

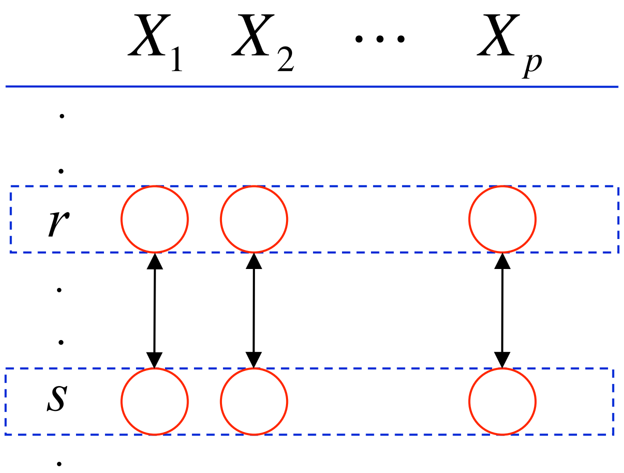
$$\begin{array}{c|c|c}
 & X_1 \quad X_2 \quad \cdots \quad X_p & \\
1 & & \\
. & & \\
r & & X \\
s & & \\
. & & \\
. & & \\
\end{array}$$

- <u>Continuous data</u>: Euclidean distance

$$D = \left(d_{ij}\right), \quad \text{where} \quad d_{rs} = \left\| o_r - o_s \right\|.$$

Define Dissimilarity

- Categorical data: Simple matching coefficient,



$$C_{rs} = \text{The proportion of features } (X_i) \text{ that are common to observations } r \text{ and } s$$

$$= \frac{(\#\ \text{of matching } X_i)}{p}$$

Define Dissimilarity

- Ordinal data: Use "ranks" as if they are continuous.

- Mixtures: (including missing data, Kaufman & Rousseeuw, 1990)

  For each feature/variable f, define first

  - If f is categorical:

$$d_{rs}^{f} = \begin{cases} 1 & \text{if } x_r^f \neq x_s^f \\ 0 & \text{otherwise} \end{cases}$$

  - If f is continuous:

$$0 \leq d_{rs}^{f} = \frac{\left| x_r^f - x_s^f \right|}{R_f} \leq 1, \text{ where } R_f \text{ is the range of } f.$$

Define Dissimilarity

Then, the dissimilarity between object r and object s is then defined as:

$$d_{rs} = \frac{\sum_f d_{rs}^f \cdot I_{rs}^f}{\sum_f I_{rs}^f}, \text{ where } I_{rs}^f = \begin{cases} 1 & \text{if } f \text{ is not missing (thus recorded)} \\ & \text{for both objects } r \text{ and } s \\ 0 & \text{otherwise} \end{cases}$$

Remark: Small score of d_{rs} indicates that objects r and s are very similar.

Properties of Dissimilarity

Dissimilarities are distance-like quantities that s.t. the following conditions for all objects *i* and *j*:

- $\delta_{ij} \geq 0$
- $\delta_{ii} = 0$
- $\delta_{ij} = \delta_{ji}$

If δ_{ij} is <u>metric</u>, then it also s.t. the triangle inequality:

$$\delta_{ij} \leq \delta_{ik} + \delta_{jk}$$

Covert Similarity to Dissimilarity

Denote by S_{ij} the "similarity" between objects *i* and *j*.

One can convert similarity to dissimilarity by using:

- $\delta_{ij} = \text{constant} - S_{ij}$
- $\delta_{ij} = 1 / S_{ij}$
- $\delta_{ij}^2 = S_{ii} + S_{jj} - 2S_{ij}$

Note that the last equality comes from:

$$\|x_i - x_j\|^2 = \|x_i\|^2 + \|x_j\|^2 - 2 < x_i, x_j >$$

➔ $< x_i, x_j >$ large (think about projection) ➔ S_{ij} large (similar)

Metric Scaling

Settings: Denote the dissimilarity matrix of the data matrix $X_{N\times p}$ by $\Delta = \{\delta_{ij}\}_{N\times N}$.

Objective: Find the best possible arrangement of the objects in a lower m-dimensional space with dissimilarity matrix

$$D = \{d_{ij}\}_{N\times N}$$

so that $\Delta \approx D$ in some appropriate norm.

Note: If δ_{ij} represents the "Euclidean distance", then we are dealing with *classical* (or Togerson-Gower) MDS.

Evaluation of D

Q: How good is the approximation $\Delta \approx D$?

➔ We employ a *loss function* and the goal is to minimize it.

(1) **Least Squares on the Distances** (Kruskal, 1964)

投影過後的data

$$\text{STRESS}(\tilde{X}) = \sum_{i=1}^{N} \sum_{j>i} w_{ij} \left(\delta_{ij} - d_{ij}(\tilde{X}) \right)^2,$$

where $\tilde{X}$ is an $N \times m$ matrix ($m < p$) that contains the coordinates of the objects in m-dimensional Euclidean space, and $d_{ij}(\tilde{X})$ denotes the distance between objects i and j in the m-dimensional space.

Notes on the STRESS Function

Notes:

- STRESS function is invariant under rotations and translations.

- STRESS function is scale dependent (e.g., not invariant under stretching and shrinking)

- A better criterion, which is not scale dependent, is the **normalized (raw) STRESS**:

$$\frac{\text{STRESS}(\tilde{X})}{\sum_{i,j} w_{ij}\delta_{ij}^2}$$

Notes on the STRESS Function

- If $w_{ij} \equiv 1$, then $\text{STRESS}(\tilde{X})$ corresponds to the Frobenius norm of $(\Delta - D)$.
- Some software packages report the square root of the **normalized stress** with $w_{ij} = 1$, called **Kruskal's Stress-1**.
- A special case of normalized stress is the so-called **Sammon mapping**, which chooses the weights as $w_{ij} = 1/\delta_{ij}$.
  定義 權重=1/相異度
  This results in

$$\textbf{Sammon's stress} = \frac{1}{\sum_{i}\sum_{j>i}\delta_{ij}} \cdot \sum_{i=1}^{N}\sum_{j>i}\frac{\left(\delta_{ij} - d_{ij}(\tilde{X})\right)^2}{\delta_{ij}}.$$

Other STRESS Functions

(2) **Least Squares on the Squared Distances**:

$$\text{STRESS}(\tilde{X}) = \sum_{i=1}^{N} \sum_{j>i} w_{ij} \left(\delta_{ij}^2 - d_{ij}^2(\tilde{X}) \right)^2 .$$

(A normalized version "SSTRESS", by Takane-Young-de Leeuw)

(3) **Least Squares on the Inner Products:** (Carroll and Chung, 1972; Meulman, 1986)

Squared distances

$$\text{STRAIN}(\tilde{X}) = \text{trace}\left\{ J\left(\Delta^2 - D^2(\tilde{X})\right) J\left(\Delta^2 - D^2(\tilde{X})\right) \right\},$$

where $\underbrace{J = I_N - {11'}/{N}}_{\text{centering operator}}$ and $1' = (1, 1, \ldots, 1)$.

Idea of Minimizing STRAIN

- To transform the dissimilarity (distance) $\delta_{ij} = \left\| \mathbf{x}_i - \mathbf{x}_j \right\|$ to inner-product $B_{ij} = <\mathbf{x}_i, \mathbf{x}_j>$.
- Recall that

$$\delta_{ij}^2 = \left\| \mathbf{x}_i - \mathbf{x}_j \right\|^2 = \left\| \mathbf{x}_i \right\|^2 + \left\| \mathbf{x}_j \right\|^2 - 2 < \mathbf{x}_i, \mathbf{x}_j >$$

- The process of "double-centering" w.r.t $\boxed{\tilde{\delta}_{ij} = -\delta_{ij}^2 / 2}$ gives:

$$\boxed{\tilde{\delta}_{ij} - \tilde{\delta}_{i\bullet} - \tilde{\delta}_{\bullet j} + \tilde{\delta}_{\bullet\bullet}} = B_{ij} = <\mathbf{x}_i, \mathbf{x}_j> .$$

- Thus, the goal is the same as minimizing:

normalization

$$\sum_{i=1}^{N} \sum_{j>i} \left(B_{ij} - < \tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j > \right)^2 \Big/ \sum_{i} \sum_{j>i} B_{ij}^2$$

Some Remarks

- A nice property of minimizing $\text{STRAIN}(\tilde{X})$ is that the dimensions of solution are nested. 要幾維度都可以, 答案會一次出來

- Finding $\tilde{X}$ using STRAIN criterion corresponds to calculating the eigen-decomposition of $-\frac{1}{2}J\Delta^{(2)}J$.

  Denote

  $$-\tfrac{1}{2}J\Delta^{(2)}J = U\Lambda U'$$

  and let m be the dimension of solution, then

  $$\tilde{X} = U_m \Lambda_m^{1/2}.$$

  The first m columns of U

  The diagonal matrix with the first m eigenvalues of Λ

Some Remarks

- Finding $\tilde{X}$ based on the STRESS criterion is non-trivial.

  The SMACOF algorithm of de Leeuw (1977) guarantees convergence to a stationary point.

- In many applications we are interested in replacing δ_{ij} by a function $f(\delta_{ij})$. For example:

$$d_{ij}(\tilde{X}) = \begin{cases} \beta \cdot \delta_{ij} + \varepsilon_{ij} & \text{(ratio scaling)} \\ \alpha + \beta \cdot \delta_{ij} + \varepsilon_{ij} & \text{(+ interval scaling)} \\ \alpha + \beta \cdot \log \delta_{ij} + \varepsilon_{ij} \\ \alpha + \beta \cdot \exp(\delta_{ij}) + \varepsilon_{ij} \end{cases}$$

➔ Finding the solution of $d_{ij}(\tilde{X})$ becomes more difficult!!

Non-metric Scaling

In some applications the dissimilarity does <u>not</u> satisfy the triangle inequality:

$$\delta_{ij} \le \delta_{ik} + \delta_{jk}$$

<u>Example</u>: (A survey form)

<u>very good</u>	<u>good</u>	<u>fair</u>	<u>poor</u>	<u>very poor</u>
5	4	3	2	1

Clearly, $5 > 3 > 2$.

But, is it true that $|5-3| > |3-2|$?

➔ This is not clear in quality!!

Non-metric Scaling

Consider a simpler constraint:

$$\delta_{ij} < \delta_{ik} \Rightarrow d_{ij}(\tilde{X}) < d_{ik}(\tilde{X})$$

➔ Such models represent only the **ordinal property** of the data.

Remark: Finding $\tilde{X}$ becomes a more challenging problem!

(see Borg & Groenen, Modern Multidimensional Scaling, Springer, 1997)

Non-metric Scaling

Solution: Consider minimizing the following version of normalized stress function

isotonic的結果 與 希望投影過後的相似度 的差距 希望最小

$$\text{STRESS} = \frac{\sum_{i,j}\left(\theta(\delta_{ij}) - d_{ij}(\tilde{X})\right)^2}{\sum_{i,j} d_{ij}^2(\tilde{X})}.$$

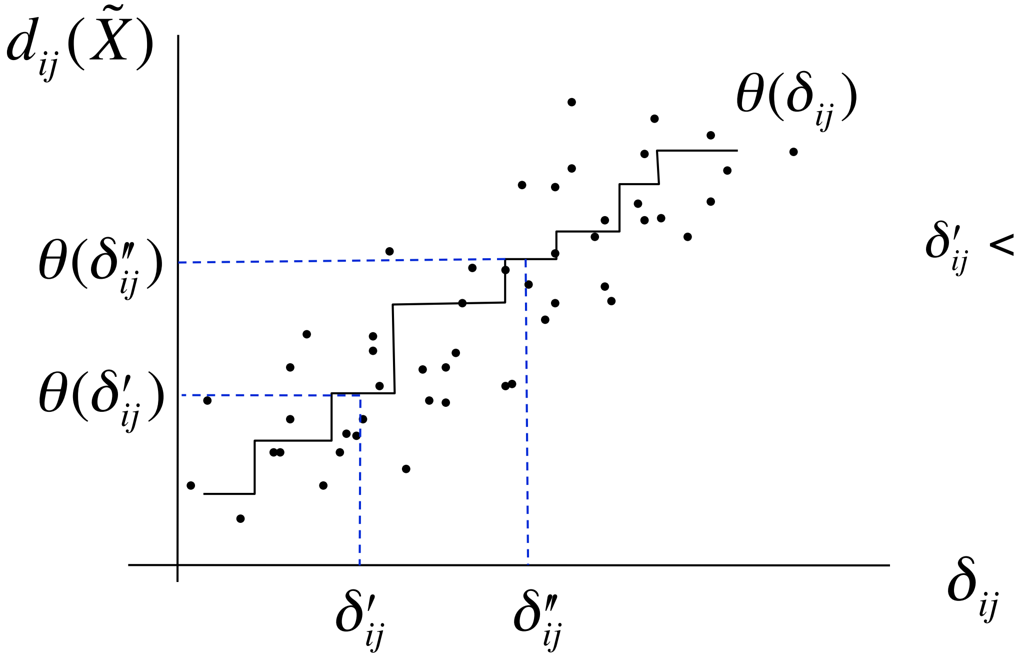
Q: How to choose $\tilde{X}$ and θ so as to minimize STRESS?

Note: Here θ is some function that preserves the property of "ordering", i.e.,

$$\delta_{ij} < \delta_{ik} \Rightarrow \theta(\delta_{ij}) \leq \theta(\delta_{ik}).$$

Utilizing Isotonic Regression

Isotonic regression is monotone regression with strictly increasing trend.



$$\delta'_{ij} < \delta''_{ij} \Rightarrow \theta(\delta'_{ij}) \le \theta(\delta''_{ij})$$
$$\Rightarrow d'_{ij}(\tilde{X}) \le d''_{ij}(\tilde{X})$$

➔ $\theta(\delta_{ij})$ is a piecewise increasing step function that minimizes the sum of squared errors (Barlow et al., 1972).

Solution to Non-metric Scaling

Algorithm

Step 1. Given $\tilde{X}$ (or $d_{ij}(\tilde{X})$), estimate $\theta(\delta_{ij})$ by using isotonic regression.

Step 2. Calculate the STRESS.

Step 3. Change $\tilde{X}$ (usually by rotations, reflections, and translations, etc), go to Step 1.

Notes:

- The best $\tilde{X}$ minimizes the STRESS.
- The algorithm can end up in local minima.

Choosing the Dimensionality of $\tilde{X}$

The usual choice of m is **2** or **3**, since we can easily plot $\tilde{X}$.

However, a rule of thumb by Kruskal (1964) is:

Sqrt(Stress)	Goodness of Fit
20%	poor
10%	fair
5%	good
2.5%	excellent
0%	perfect

In practice, we can pick up a large enough m so that the fit is at least "fair".

An Illustrative Example

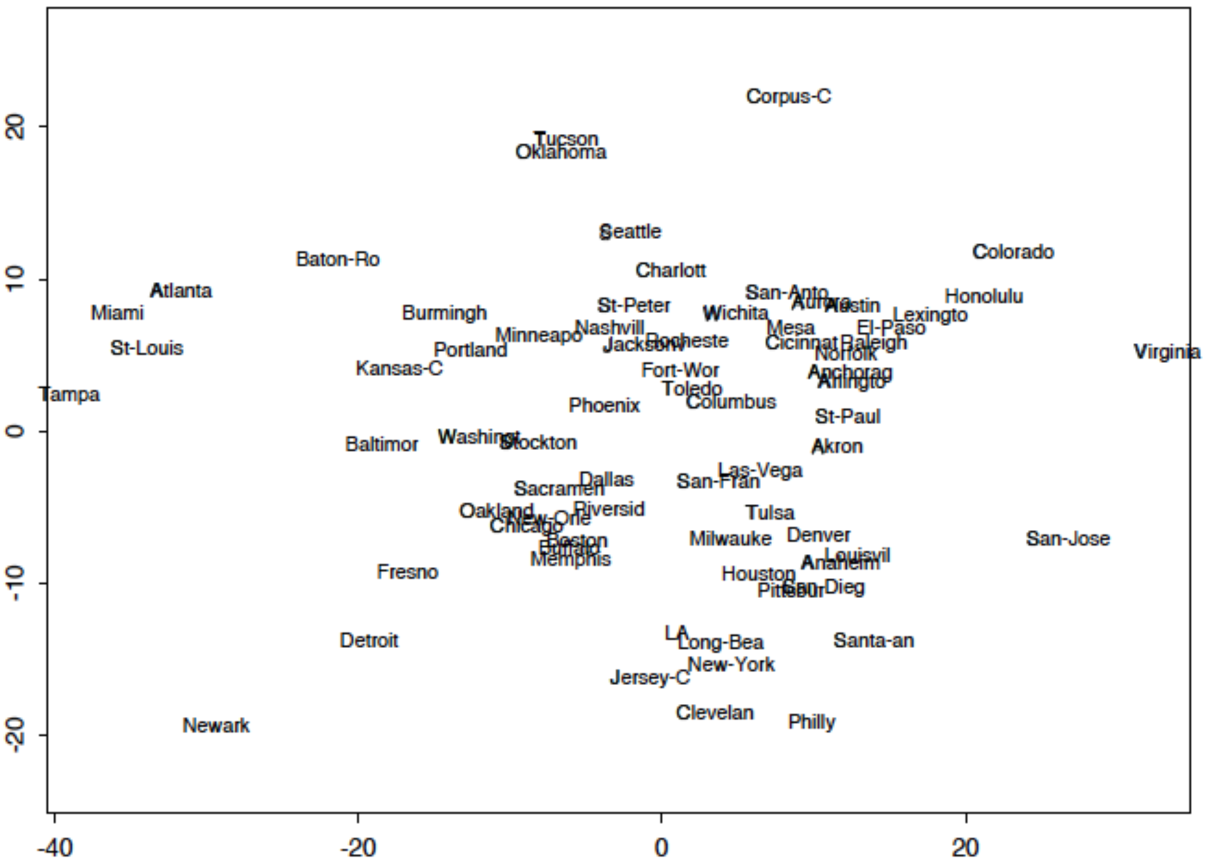


Fig. A 2D metric MDS solution of City Crime Data

Note: The 2D representation is very similar to that of PCA.