Multivariate Data Analysis

(MGT513, BAT531, TIM711)

Lecture 5

Multidimensional Scaling (MDS)

References

1. Analyzing Multivariate Data (LCG)

[Ch 7: Multidimensional Scaling]

2. Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning by Alan Julian Izenman

[Ch 13: Multidimensional Scaling and Distance Geometry]

❖ E-book is available at the UNIST library

Introduction

- We can talk about the proximity of any two entities to each other, whereby “entity” we might mean an object, a brand-name product, a nation, a stimulus, etc.
- The proximity of a pair of such entities could be a measure of association(e.g., the absolute value correlation coefficient), a confusion frequency (i.e., to what extent one entity is confused with another in an identification exercise), or some other measure of how alike (or different) one perceives the entities.

Introduction

- The general problem of multidimensional scaling (MDS) essentially reverses that relationship.
- Given only a two-way table of proximities, we wish to reconstruct the original map as closely as possible (We do not know the number of dimensions in which the entities are located.)
- MDS is a family of algorithms, each designed to arrive at an optimal low-dimensional configuration for a particular type of proximity data.
- MDS is a primarily a data visualization method for identifying “clusters” of points (when the dimension is low).

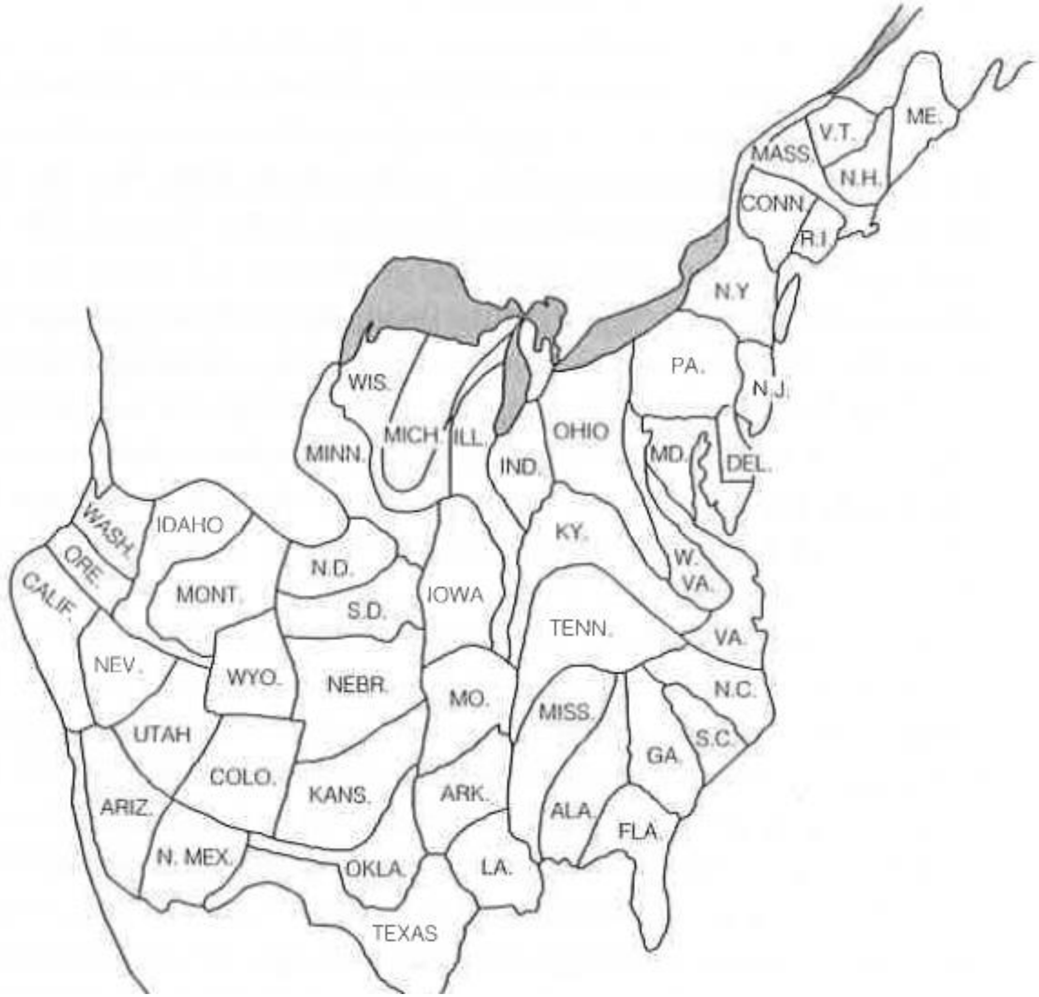
Multidimensional Scaling (MDS)

- Goal of Multidimensional scaling (MDS): Given pairwise dissimilarities(δ_{ij}) (NOT need to be a metric), reconstruct a map that preserves distances.
- MDS is a family of different algorithms that attempt to find optimal low-dimensional configuration, say t = 2 or t = 3
- Reconstructed map has coordinates $\mathbf{X}_i = (X_{i1}, X_{i2} \ldots, X_{it})$ and the *Euclidean* distance $\| \mathbf{X}_i - \mathbf{X}_j \|$
- MDS methods include

1. Classical MDS [uses eigen-decomposition]
2. Distance Scaling [uses iterative procedures]
   – Metric MDS
   – Non-metric MDS

Application: Perceptual Mapping

FIGURE 7.1
Shepard's "Bostonian's View of the United States"



Application: Perceptual Mapping

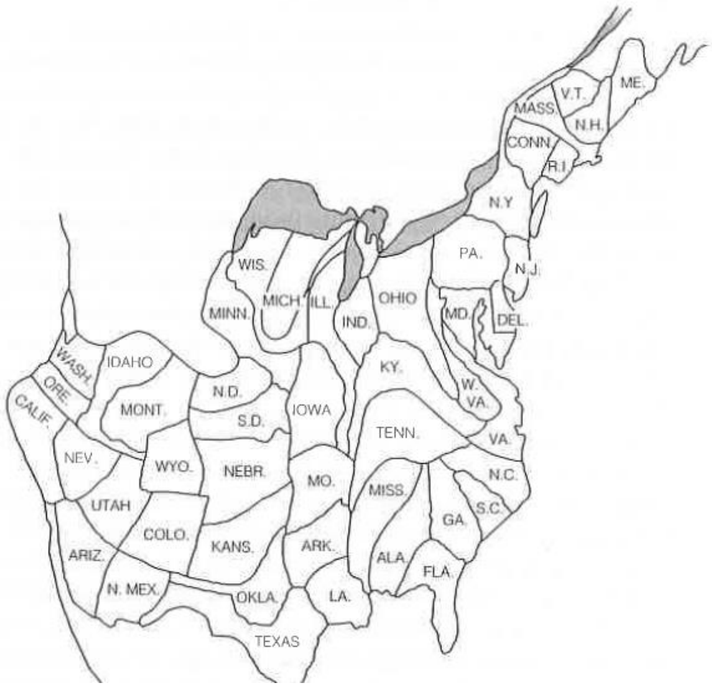
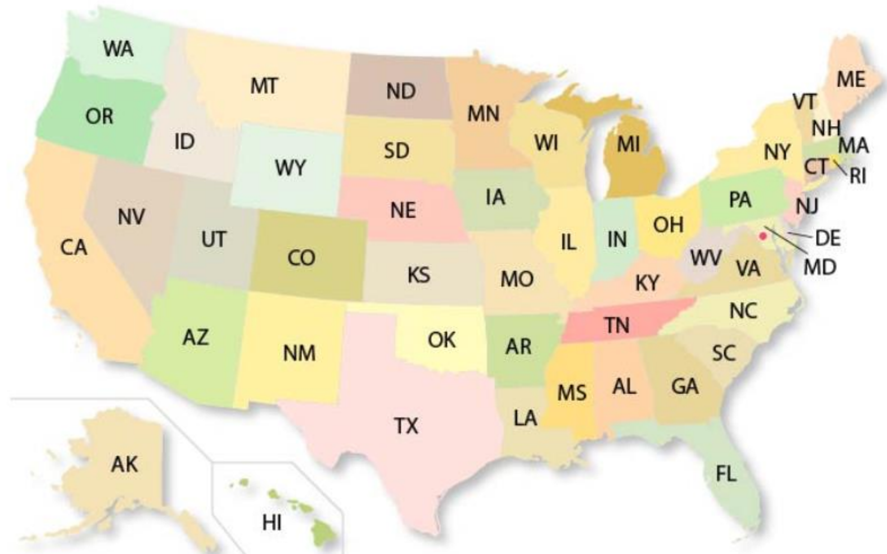


FIGURE 7.1 Shepard's "Bostonian's View of the United States"



Source: https://www.nationsonline.org/

Distance Measures

- Manhattan distance (city-block distance)
  - Sum of the absolute differences of the variables
  - Minkowski distance of order 1 or L_1 norm
  - $d_{ij} = \sum_{k=1}^{n} |x_{ik} - x_{jk}|$
- Euclidean distance
  - Straight-line distance
  - Minkowski distance of order 2 or L_2 norm
  - $d_{ij} = (\sum_{k=1}^{n} |x_{ik} - x_{jk}|^2)^{1/2}$
- Maximum distance (Chebyshev distance)
  - The largest distance on any one dimensions
  - $d_{ij} = \max_k(|x_{ik} - x_{jk}|)$

Distance Measures

- Minkowski distance
  - L_p norm distance
  - The p th root of the sum of the p th powers of the differences of the components
  - $d_{ij} = (\sum_{k=1}^{n} |x_{ik} - x_{jk}|^{\mathrm{p}})^{1/p}$
  - Manhattan distance (p=1), Euclidean distance (p=2), Maximum distance (p=∞)
- Mahalanobis distance
  - Generalized Euclidean distance which accounts for the standard deviation of each variable and the correlations among variables
  - $d_{ij} = [(x_i - x_j)^T C^{-1}(x_i - x_j)]^{1/2}$ where C is the sample covariance matrix

Proximity Matrices

- The focus on pairwise comparisons of entities is fundamental to MDS.
- The “closeness” of two entities are measured by a proximity measure.
- Proximity can be a continuous measure of how physically close one entity is to another or it could be a subjective judgment recorded on an ordinal scale
- Proximity data are obtained from a group of subjects, each of whom make similarity (or dissimilarity) judgements on all possible m unordered pairs of n entities.

$$m = \binom{n}{2} = \frac{1}{2}n(n-1)$$

Proximity Matrices

- Consider a particular collection of n entities.
- Let δ_{ij} represent the dissimilarity of the ith entity to the jth entity
- We arrange the n dissimilarities, $\{\delta_{ij}\}$, into $(n \times n)$ square matrix,

$$\boldsymbol{\Delta} = (\delta_{ij})$$

called a **proximity matrix**, where

$$\delta_{ij} \geq 0, \qquad \delta_{ii} = 0, \qquad \delta_{ji} = \delta_{ij}$$

Classical MDS

- Suppose we are given n points $\mathrm{Y}_1, \ldots, \mathrm{Y}_n \in \Re^r$.
- From these points, we can compute an $(n \times n)$ proximity matrix $\boldsymbol{\Delta} = (\delta_{ij})$ of dissimilarities (***<u>Euclidean distances</u>***) between the points $\mathrm{Y}_i = (Y_{ik})$ and $\mathrm{Y}_j = (Y_{jk})$.

$$\delta_{ij} = \| \mathrm{Y}_i - \mathrm{Y}_j \| = \left\{ \sum_{k=1}^{r} (Y_{ik} - Y_{jk})^2 \right\}^{1/2}$$

- [Note: distances other than *Euclidean* (e.g. Minkowski L_p distance) can be employed for non-classical MDS (metric MDS)].

Classical MDS

- In the dimensionality-reduction aspect of MDS, we wish to find a t dimensional representation, $\mathrm{X}_1, \ldots, \mathrm{X}_n \in \Re^t$(referred to as principal coordinates), of those r-dimensional points (with $t < r$), such that the interpoint distances in t-space "match" those in r-space.

$$\delta_{ij} = \| \mathrm{Y}_i - \mathrm{Y}_j \| \approx \| \mathrm{X}_i - \mathrm{X}_j \|$$

- When dissimilarities (δ_{ij}) are defined as *Euclidean* interpoint distances, this type of "classical" is equivalent to PCA in that the principal coordinates are identical to the scores of the first t principal components of the $\{\mathrm{Y}_i\}$.

Classical MDS

- Typically, in classical scaling (Torgerson, 1952, 1958), $\{\mathbf{Y}_i\} \subset \Re^r$ are not given.
- Instead, we are given only the (*Euclidean*) dissimilarities $\{\delta_{ij}\}$ through the proximity matrix $\boldsymbol{\Delta}$.
- The objective of classical Multidimensional Scaling (cMDS) is to find a configuration $\mathbf{X} = [\mathrm{X}_1, \dots, \mathrm{X}_n]$ so that $\| \mathbf{X}_i - \mathbf{X}_j \| = \delta_{ij}$.
- Such a solution is *not unique*: if $\mathbf{X}$ is the solution, then $\mathbf{X}^* = \mathbf{X} + c, c \in \Re^r$ is also a solution.
- Therefore, the configuration need to be centered:

$$\sum_{i=1}^{n} X_{ik} = 0, \qquad \text{for } k = 1,2,..,r$$

Classical MDS

- The Euclidean distance between the i-th and the j-th point is:

$$\| \mathrm{X}_i - \mathrm{X}_j \|^2 = \mathrm{X}_i{}^T \mathrm{X}_i + \mathrm{X}_j{}^T \mathrm{X}_j - 2\mathrm{X}_i{}^T \mathrm{X}_j$$

- We use the $(n \times n)$ Gram matrix $\boldsymbol{B} = \mathbf{X}^T\mathbf{X}$ rather than $\mathbf{X}$. Then b_{ij} term of $\boldsymbol{B}$ is given by:

$$b_{ij} = \mathrm{X}_i{}^T \mathrm{X}_j$$

- We derive $\boldsymbol{B}$ from the *known* squared distances δ_{ij}, and then derive unknown coordinates $\mathbf{X}$ from $\boldsymbol{B}$:

$$\delta_{ij}{}^2 = \mathrm{X}_i{}^T \mathrm{X}_i + \mathrm{X}_j{}^T \mathrm{X}_j - 2\mathrm{X}_i{}^T \mathrm{X}_j = b_{ii} + b_{jj} - 2b_{ij} \quad (1)$$

Classical MDS

- Due to centering of the coordinate matrix $\mathbf{X}$, $\sum_{i=1}^{n} b_{ij} = 0$. Summing (1) over i, over j, and over i and j:

$$\frac{1}{n}\sum_{i=1}^{n} {\delta_{ij}}^2 = \frac{1}{n}\sum_{i=1}^{n} b_{ii} + b_{ij}$$

$$\frac{1}{n}\sum_{j=1}^{n} {\delta_{ij}}^2 = b_{ii} + \frac{1}{n}\sum_{j=1}^{n} b_{jj}$$

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} {\delta_{ij}}^2 = \frac{2}{n}\sum_{i=1}^{n} b_{ii}$$

(2)

Classical MDS

- Combining (1) and (2), the solution is unique:

$$b_{ij} = -\frac{1}{2}({\delta_{ij}}^2 - {\delta_{i.}}^2 - {\delta_{.j}}^2 + {\delta_{..}}^2)$$

- A solution $\mathbf{X}$ is given by the eigen-decomposition of $\boldsymbol{B}$.

  For $\boldsymbol{B} = \boldsymbol{V\Lambda V^T}$,

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

  where $\boldsymbol{\Lambda} = \{\boldsymbol{\lambda_1}, \boldsymbol{\lambda_2}, .., \boldsymbol{\lambda_n}\}$ and $\boldsymbol{V} = [\boldsymbol{v_1}, \boldsymbol{v_2}, .., \boldsymbol{v_n}]$

Classical MDS

- We may choose the first t rows of $\mathbf{X}$ which best preserves the distances δ_{ij} among all other linear dimension reduction of $\mathbf{X}$ (to dimension t).

$$\mathbf{X}_{\boldsymbol{t}} = \boldsymbol{\Lambda}_{\boldsymbol{t}}^{1/2} \boldsymbol{V}_{\boldsymbol{t}}^{\boldsymbol{T}}$$

  where $\boldsymbol{\Lambda}_t$ is the first $t \times t$ sub-matrix of $\boldsymbol{\Lambda}$ and $\boldsymbol{V}_{\boldsymbol{t}}$ is the first t column of $\boldsymbol{V}$.

- cMDS offers configurations $[\mathrm{X}_{1(t)}, \dots, \mathrm{X}_{n(t)}]$ where each columns belong to $\Re^t$
- cMDS on Euclidean distance is equivalent to PCA
- It is also called as Principal Coordinate Analysis (PCoA).

cMDS Example

TABLE 7.1 Distance between cities in Europe (in miles, "as the crow flies")

	Athens	Berlin	Dublin	London	Madrid	Paris	Rome	Warsaw
Athens								
Berlin	1,119							
Dublin	1,777	817						
London	1,486	577	291					
Madrid	1,475	1,159	906	783				
Paris	1,303	545	489	213	652			
Rome	646	736	1,182	897	856	694		
Warsaw	1,013	327	1,135	904	1,483	859	839	

cMDS Example

TABLE 7.2 Matrix **B** for European cities data

1076014	−12160	−770338	−467392	−238364	−273582	438858	246968
−12160	151827	12688	8148	−284286	−35284	−85426	244495
−770338	12688	541039	326878	171543	188274	−318534	−151547
−467392	8148	326878	197399	103597	113331	−194096	−87863
−238364	−284286	171543	103597	622883	136205	54582	−566157
−273582	−35284	188274	113331	136205	74631	−93994	−109579
438858	−85426	−318534	−194096	54582	−93994	219018	−20406
246968	244495	−151547	−87863	−566157	−109579	−20406	444092

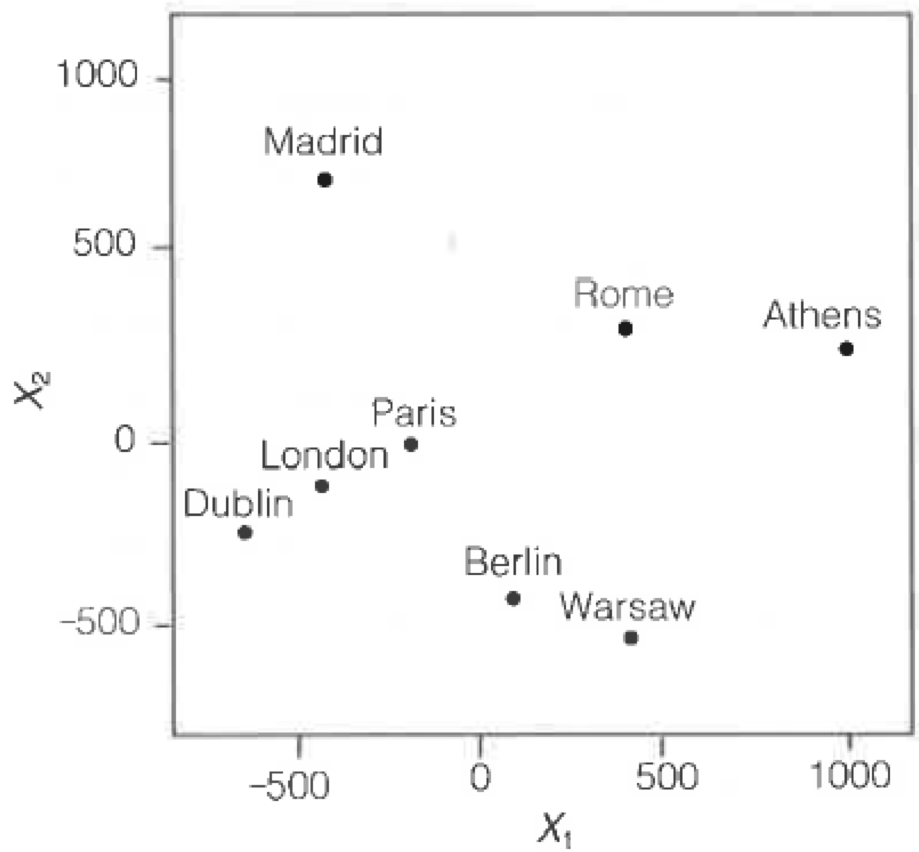
cMDS Example

TABLE 7.3 Singular value decomposition of matrix **B** for European cities data (first two dimensions)

	Eigenvalue	X_1	X_2
1	2240139	1011	239
2	1131445	77	−375
3	−54323	−715	−184
4	11084	−432	−114
5	−1652	−407	688
6	250	−274	28
7	−46	368	290
8	3	372	−573

cMDS Example

FIGURE 7.5
Classical metric MDS solution for European city data



Distance Scaling

- **Classical MDS**

  We wish to find a configuration of points in a lower-dimensional space such that

$$\delta_{ij} \approx d_{ij} = \| \mathrm{X}_i - \mathrm{X}_j \|$$

- **Distance Scaling**

  In distance scaling, this relationship is relaxed. We wish to find a suitable configuration such that

$$f(\delta_{ij}) \approx d_{ij} = \| \mathrm{X}_i - \mathrm{X}_j \|$$

where f is some *monotonic* function which transforms the dissimilarities into distances.

Distance Scaling

- Two Types of Distance Scaling
  1. Metric distance scaling (metric MDS):
     if dissimilarities δ_{ij} are quantitative
  1. Non-metric distance scaling (non-metric MDS):
     if dissimilarities δ_{ij} are qualitative (e.g. ordinal)

- Unlike cMDS, distance scaling is an optimization process minimizing *stress* function, and is solved by iterative algorithms.

Metric MDS

- Dissimilarities δ_{ij} are quantitative measurements.
- A function $\delta(u_i, u_j) = \delta_{ij}$ is a dissimilarity function. In mathematics, a distance function (that gives a distance between two objects) is also called ***metric***, satisfying
  1. $\delta_{ij} \geq 0$ **[non-negative]**
  2. $\delta_{ij} = 0$ if and only if $u_i = u_j$ **[identity property]**
  3. $\delta_{ij} = \delta_{ji}$ **[symmetric]**
  4. $\delta_{ij} \leq \delta_{ik} + \delta_{kj}$ **[triangle inequality]**
- Distances other than *Euclidean* can be employed for Metric MDS.

Metric MDS

- The function f is usually taken to be a *parametric monotonic function*, such as $f(\delta_{ij}) = \alpha + \beta\delta_{ij}$

  where α and β are unknown positive coefficients.

  1. absolute MDS: if $\alpha = 0,\ \beta = 1$
  2. ratio MDS: if $\alpha = 0,\ \beta > 1$
  3. interval MDS: if $\alpha \geq 0,\ \beta \geq 1$

- If the $\{\delta_{ij}\}$ are similarities (rather than dissimilarity), then we need $\beta < 0$.

- Useful reference: http://cda.psych.uiuc.edu/mds_509_2013/borg_groenen/chapter_nine.pdf

Metric MDS (Metric Least-Squares Scaling)

- The distances $\{d_{ij}\}$ can be fitted to $\{f(\delta_{ij})\}$ by least-squares (LS). The result is metric LS scaling.
- A given configuration of points $\{X_{ij}\} \subset \Re^t$ can be evaluated by a *loss function*,

$$\mathcal{L}_f(\mathrm{X}_{1,\dots,}\mathrm{X}_n; \mathbf{W}) = \sum_{i<j} w_{ij} \left(d_{ij} - f(\delta_{ij})\right)^2$$

where $\mathbf{W} = (w_{ij})$ is a given matrix of weights.

Metric MDS
(Metric Least-Squares Scaling)

- Metric stress function:

$$\text{stress} = [\mathcal{L}_f(\mathrm{X}_{1,\ldots,}\mathrm{X}_n; \mathbf{W})]^{1/2}$$

- Minimizing stress over all t-dimensional configurations $\{X_{ij}\}$ and monotone f yields an optimal metric distance scaling solution.

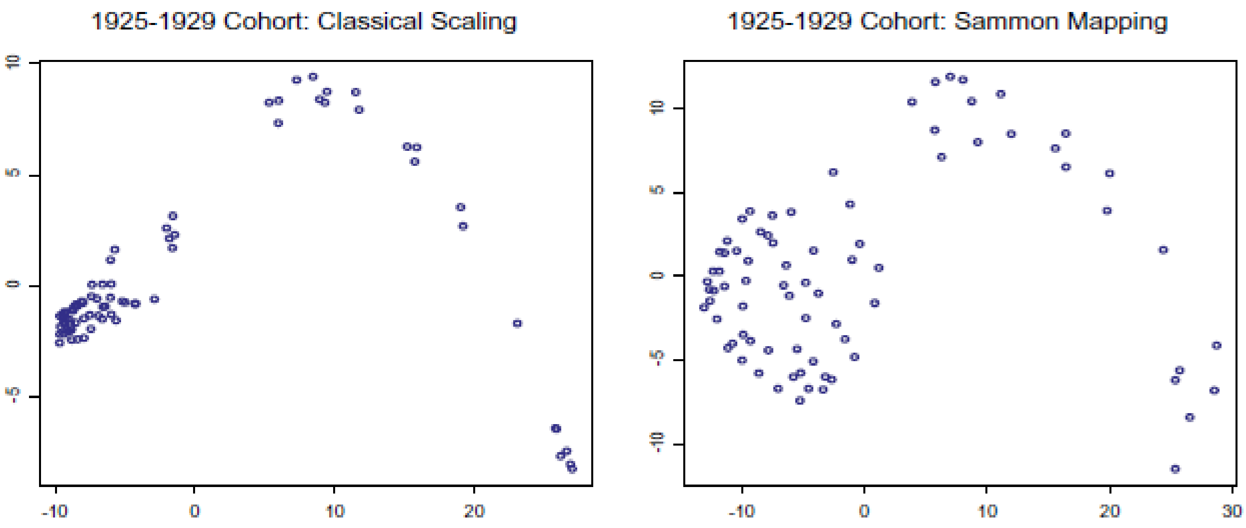
- Weighting systems include
  - $w_{ij} = \{\sum_{k<l} \delta_{kl}{}^2\}^{-1}$
  - $w_{ij} = \delta_{ij}{}^{-2}$
  - $w_{ij} = \delta_{ij}{}^{-1}\{\sum_{k<l} \delta_{kl}\}^{-1}$: *Sammon Mapping* (nonlinear least-squares scaling)

- Useful reference: https://www.jstatsoft.org/article/view/v073i08

UNIST ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY 2009

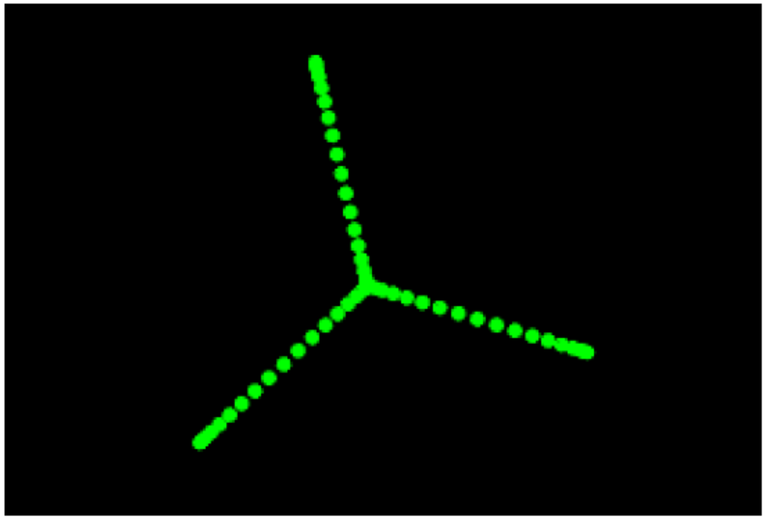
Metric MDS
Sammon Mapping

- Sammon mapping can preserve the small dissimilarities by giving them a greater degree of importance than larger dissimilarities.

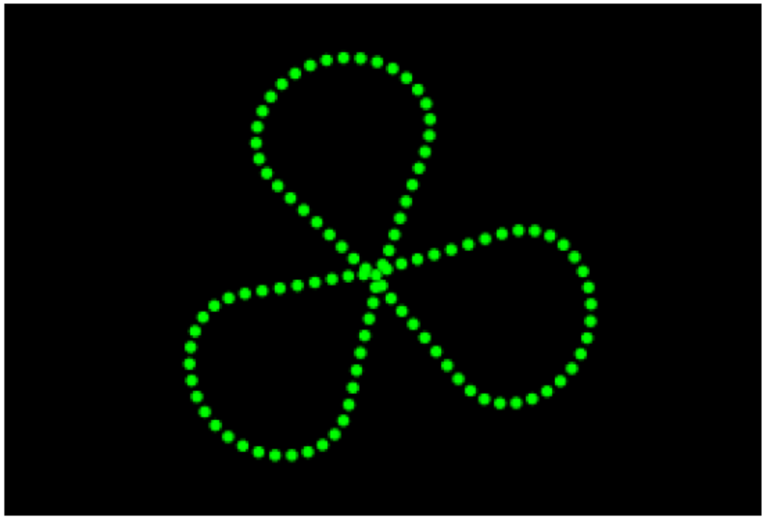


Source: Izenman Figure 13.9 (lower panel)

Metric MDS
Sammon Mapping



(a) Projection by PCA does not preserve the structure of the dataset — it is unclear that it consists of three circles



(b) The Sammon mapping preserves the topological structure — while the circles become distorted, there are still three closed loops meeting at a single point

Figure 2: PCA and Sammon projections of a six-dimensional 'bouquet of circles', from [3]. The original dataset contains three mutually perpendicular circles in six-dimensional space, meeting at a point

Source: P. Henderson "Sammon Mapping."https://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/AV0910/henderson.pdf

Non-metric MDS

- Main idea: Uncovering metric structures from ordinal data on similarities
- In many applications of MDS, dissimilarities are known only by their rank order, and the spacing between successively ranked dissimilarities is of no interest or is unavailable.
- Objective of non-metric MDS (by Roger Shepard)
  - To achieve a monotone relationship between the observed dissimilarities (δ_{ij}) and the fitted distances in the scaling configuration (d_{ij}).

Non-metric MDS: ordinal dissimilarity

TABLE 7.4 Perceived dissimilarities for different car models

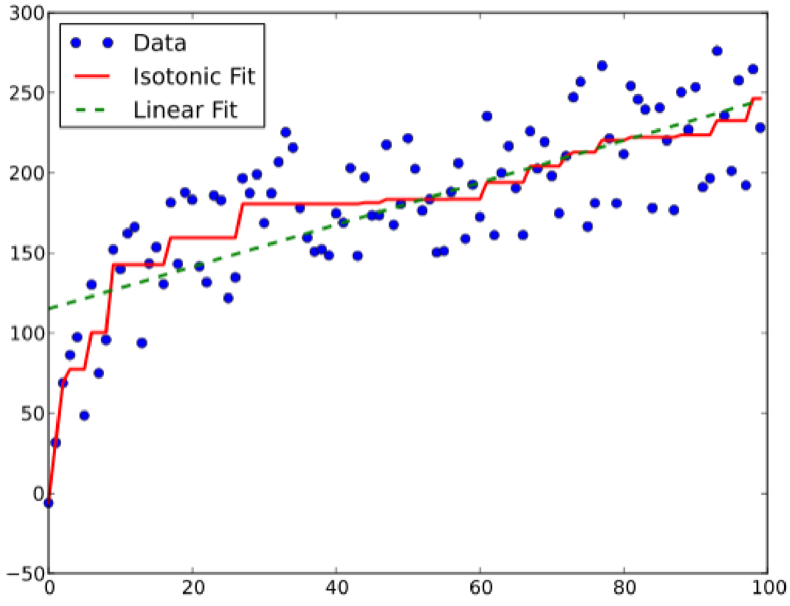
	BMW	Ford	Infiniti	Jeep	Lexus	Chrysler	Mercedes	Saab	Porsche	Volvo
BMW										
Ford	34									
Infiniti	8	24								
Jeep	31	2	25							
Lexus	7	26	1	27						
Chrysler	43	14	35	15	37					
Mercedes	3	28	5	29	4	42				
Saab	10	18	20	17	13	36	19			
Porsche	6	39	41	38	40	45	32	21		
Volvo	33	11	22	12	23	9	30	16	44	

Non-metric MDS

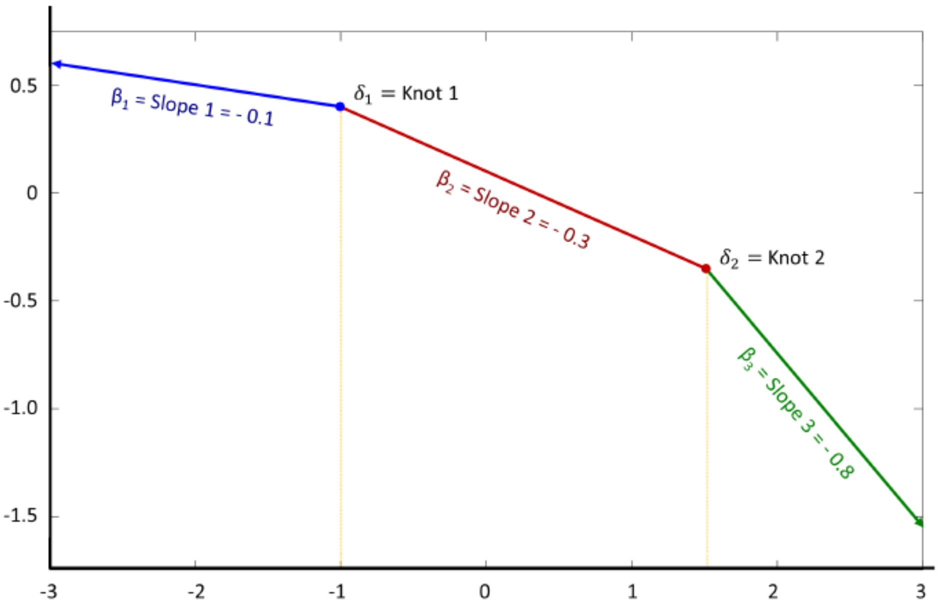
- Given a (low) dimension *t*, non-metric MDS seeks to find an optimal configuration $\mathbf{X} \subset \Re^t$ that gives
$$f(\delta_{ij}) \approx d_{ij} = \| \mathrm{X}_i - \mathrm{X}_j \|$$
- In non-metric MDS, *f* is much general than metric MDS and is only *implicitly* defined.
- *Disparities* $\hat{d}_{ij} = f(\delta_{ij})$. It only preserved the rank order of δ_{ij}:
$$\delta_{i_1 j_1} < \delta_{i_2 j_2} < \cdots < \delta_{i_m j_m}$$
$$\Leftrightarrow \hat{d}_{i_1 j_1} < \hat{d}_{i_2 j_2} < \cdots < \hat{d}_{i_m j_m}$$
$$\Leftrightarrow d_{i_1 j_1} < d_{i_2 j_2} < \cdots < d_{i_m j_m}$$
- f: Computing (nondecreasing) disparities: isotonic regression, monotone spline

Non-metric MDS

Isotonic Regression



Spline Regression



Source: Wiki

Non-metric MDS

- Kruskal's non-metric MDS stress-1 (badness of fit):

$$\text{stress}-1 = \left\{ \frac{\sum_{i<j} (d_{ij} - \hat{d}_{ij})^2}{\sum_{i<j} d_{ij}^2} \right\}^{1/2}$$

- Function f works as if it were a regression curve (approximated dissimilarities d_{ij} as y, disparities $\hat{d}_{ij}$ as $\hat{y}$, and the rank order of dissimilarities (δ_{ij}) as explanatory) – **[Shepard diagram]**
- Repeat (given a (low) dimension *t*)
  - Change the configuration of points by applying an iterative gradient search algorithm (e.g., method of steepest descent) to the stress criterion. This step will produce a new set of $\{d_{ij}\}$.
  - Use an isotonic regression algorithm to produce revised values of the $\{\hat{d}_{ij}\}$, together with a smaller stress value.

UNIST ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY 2009

Non-metric MDS

- Shepard diagram (Isotonic regression / Monotone spline)

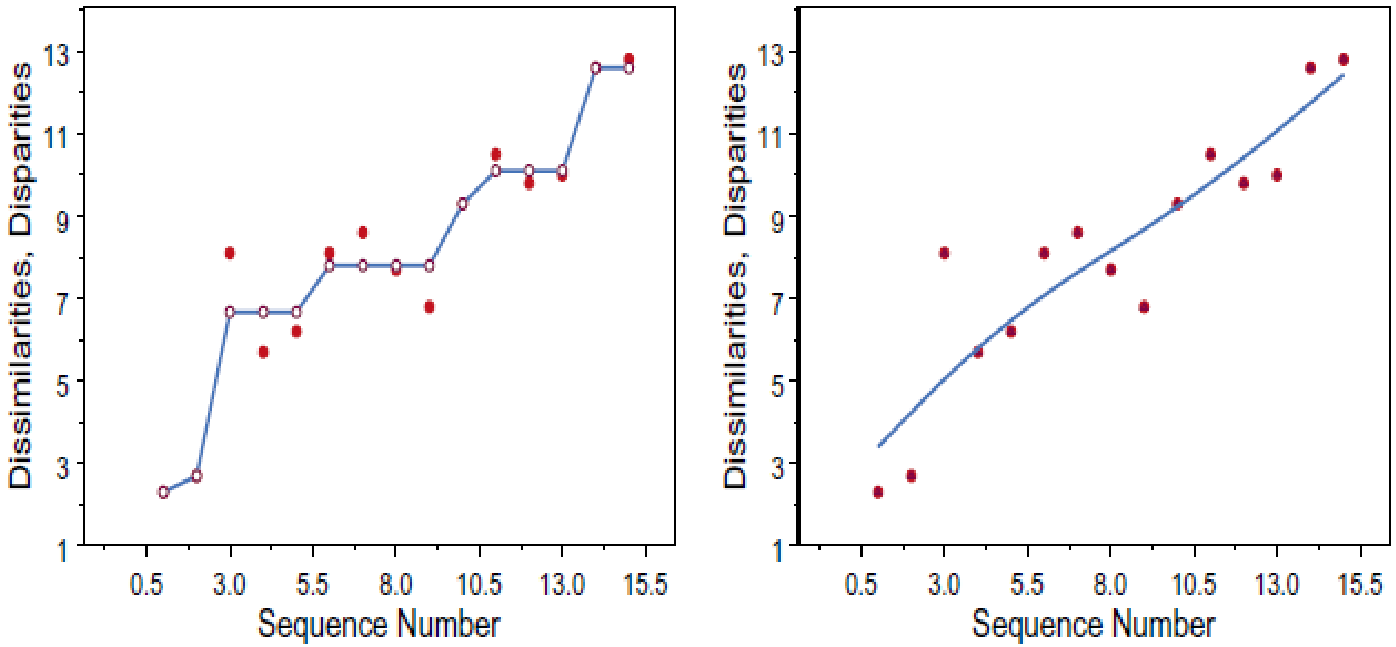


FIGURE 13.10. *Shepard diagram for the artificial example. Left panel: Isotonic regression. Right panel: Monotone spline. Horizontal axis is rank order. For the red points, the vertical axis is the dissimilarity d_{ij}, whereas for the fitted blue points, the vertical axis is the disparity $\widehat{d}_{ij}$.*

Non-metric MDS

TABLE 13.9. *The nonmetric distance-scaling algorithm.*

1. Order the $m = \frac{1}{2}n(n-1)$ dissimilarities $\{\delta_{ij}\}$ from smallest to largest as in (13.25).
2. Fix the number t of dimensions and choose an initial configuration of points $\mathbf{y}_i \in \Re^t$, $i = 1, 2, \ldots, n$.
3. Compute the set of distances $\{d_{ij}\}$ between all pairs of points in the initial configuration.
4. Use an isotonic regression algorithm to produce fitted values $\{\widehat{d_{ij}}\}$. Compute the initial value of stress.
5. Change the configuration of points by applying an iterative gradient search algorithm (e.g., method of steepest descent) to the stress criterion. This step will produce a new set of $\{d_{ij}\}$.
6. Use an isotonic regression algorithm to produce revised values of the $\{\widehat{d_{ij}}\}$, together with a smaller stress value.
7. Repeat steps 5 and 6 until the current configuration produces a minimum stress value, so that no further improvement in stress can take place by further reconfiguring the points.
8. Repeat the previous steps using a different value of t. Plot stress against t. Choose that value of t that gives a reasonably small value of stress and where no significant decrease in stress can result from increasing t. This is usually exhibited by an "elbow" in the plot.

Non-metric MDS

- **Goodness of fit**

TABLE 13.10 *Evaluation of "stress."*

Stress	Goodness of Fit
0.20	Poor
0.10	Fair
0.05	Good
0.025	Excellent
0.0	"Perfect"

- **Optimal dimension (t)**

  Draw the scree plot of stress and compare stress values from different dimensions. Choose that value of t for which the minimum stress is small and any further increase in t does not significantly decrease the minimum stress.

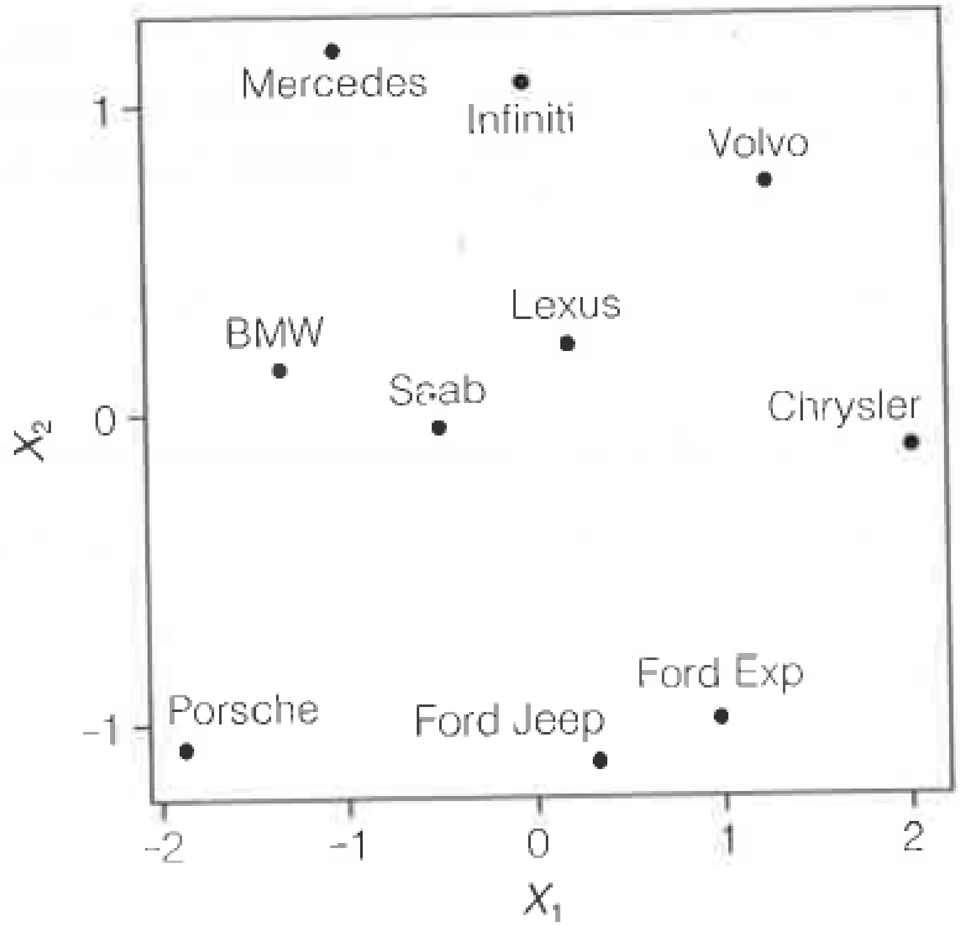
Non-metric MDS Example

TABLE 7.4 Perceived dissimilarities for different car models

	BMW	Ford	Infiniti	Jeep	Lexus	Chrysler	Mercedes	Saab	Porsche	Volvo
BMW										
Ford	34									
Infiniti	8	24								
Jeep	31	2	25							
Lexus	7	26	1	27						
Chrysler	43	14	35	15	37					
Mercedes	3	28	5	29	4	42				
Saab	10	18	20	17	13	36	19			
Porsche	6	39	41	38	40	45	32	21		
Volvo	33	11	22	12	23	9	30	16	44	

Non-metric MDS Example

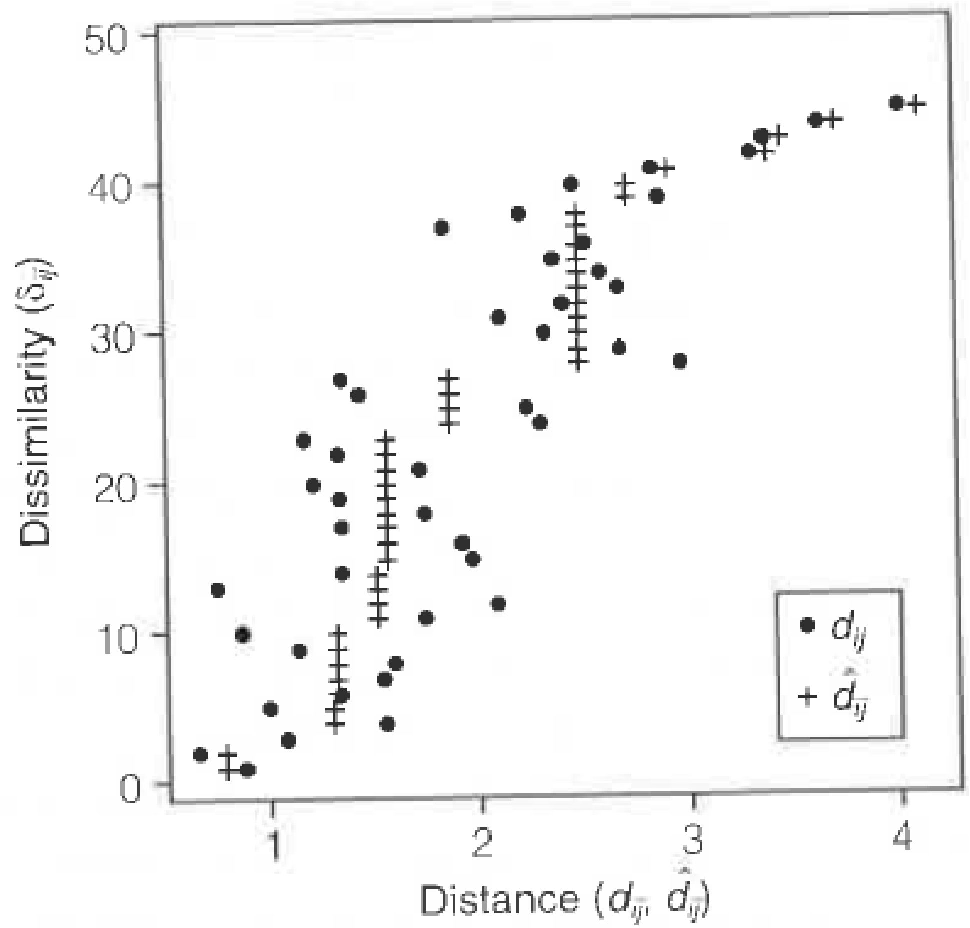
FIGURE 7.6
Arbitrary initial configuration for car data



Non-metric MDS Example

FIGURE 7.7
Shepard diagram for initial configuration for car data

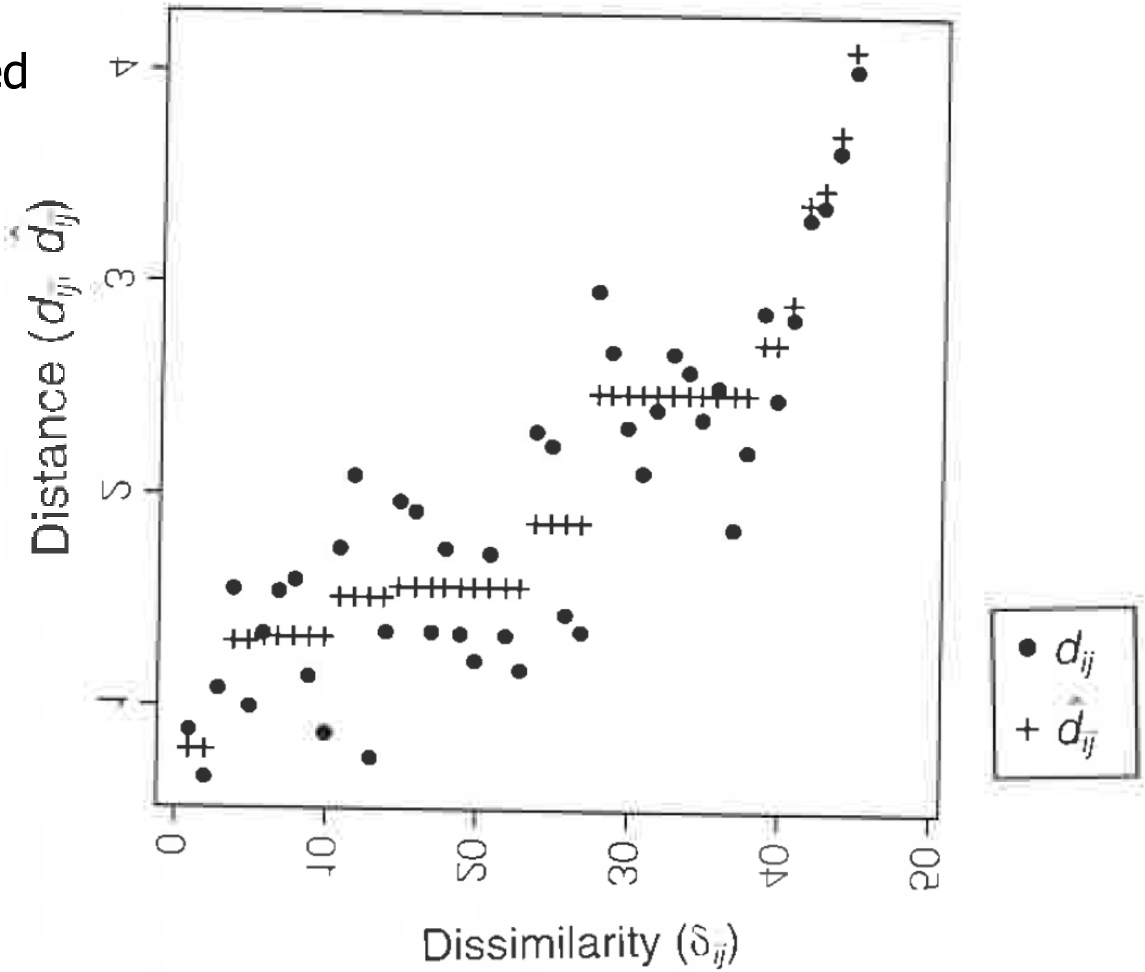
Stress = 0.14



Non-metric MDS Example

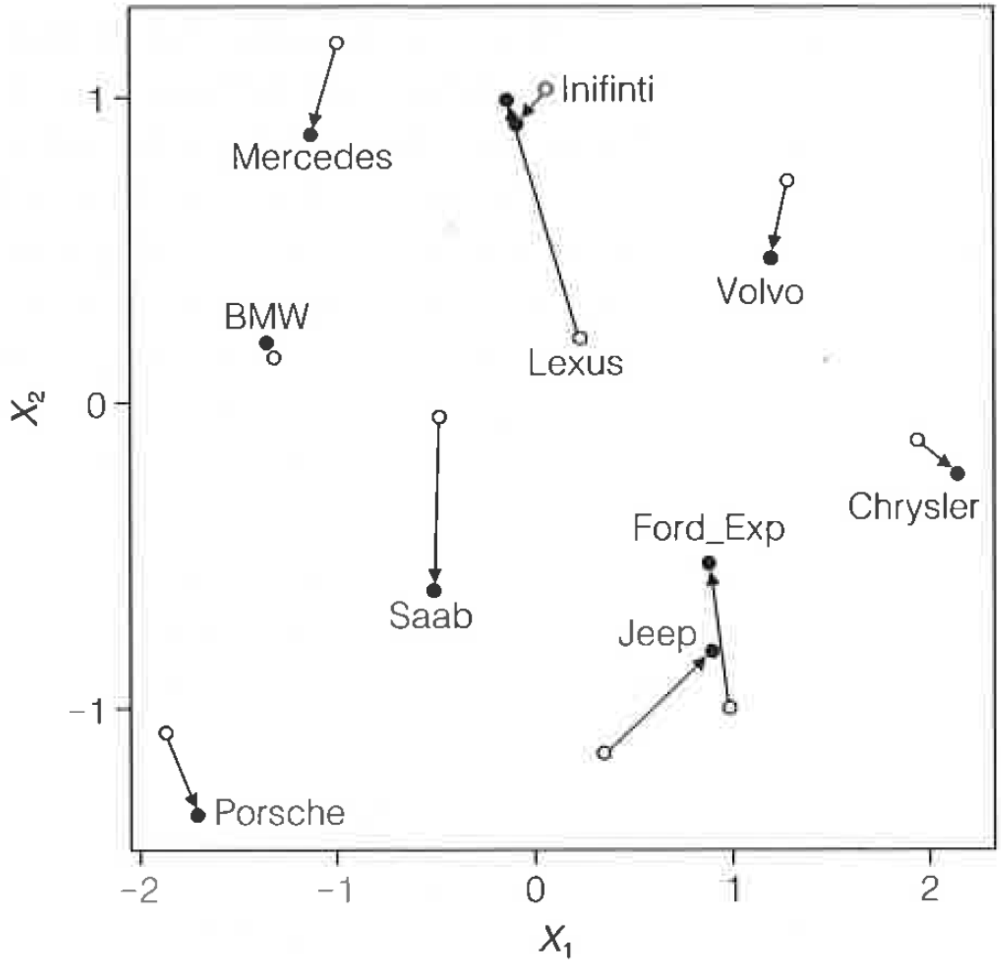
Reflected & Rotated

Stress = 0.14



Non-metric MDS Example

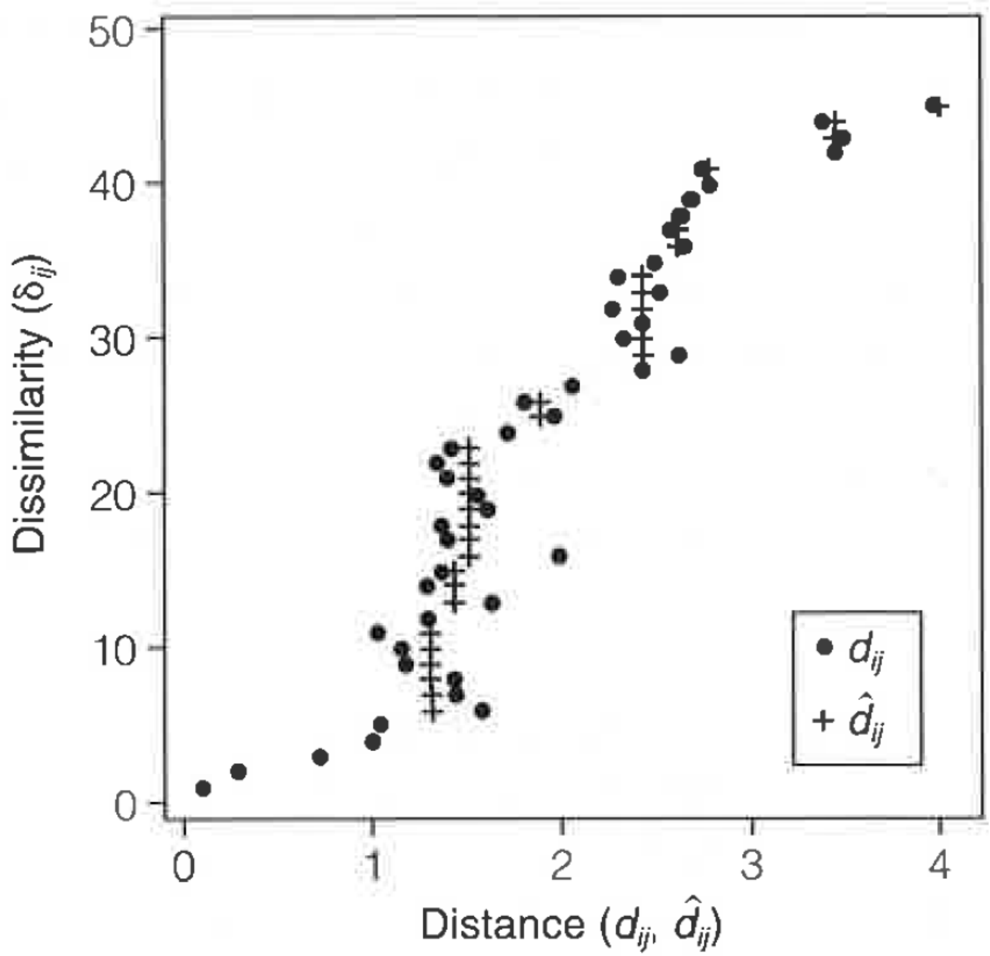
FIGURE 7.8
Plot showing change after one iteration



Non-metric MDS Example

FIGURE 7.9
Shepard diagram showing improvement after iteration

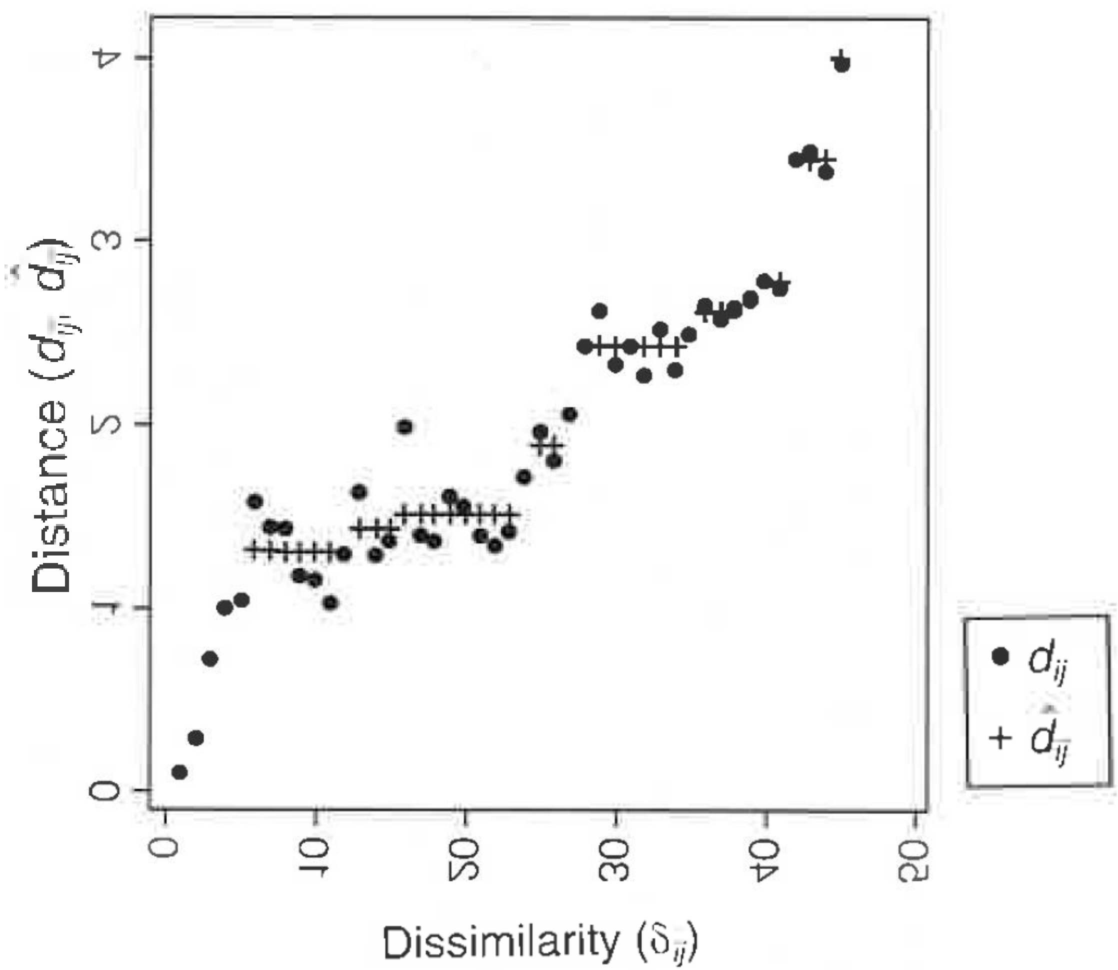
Stress = 0.06



Non-metric MDS Example

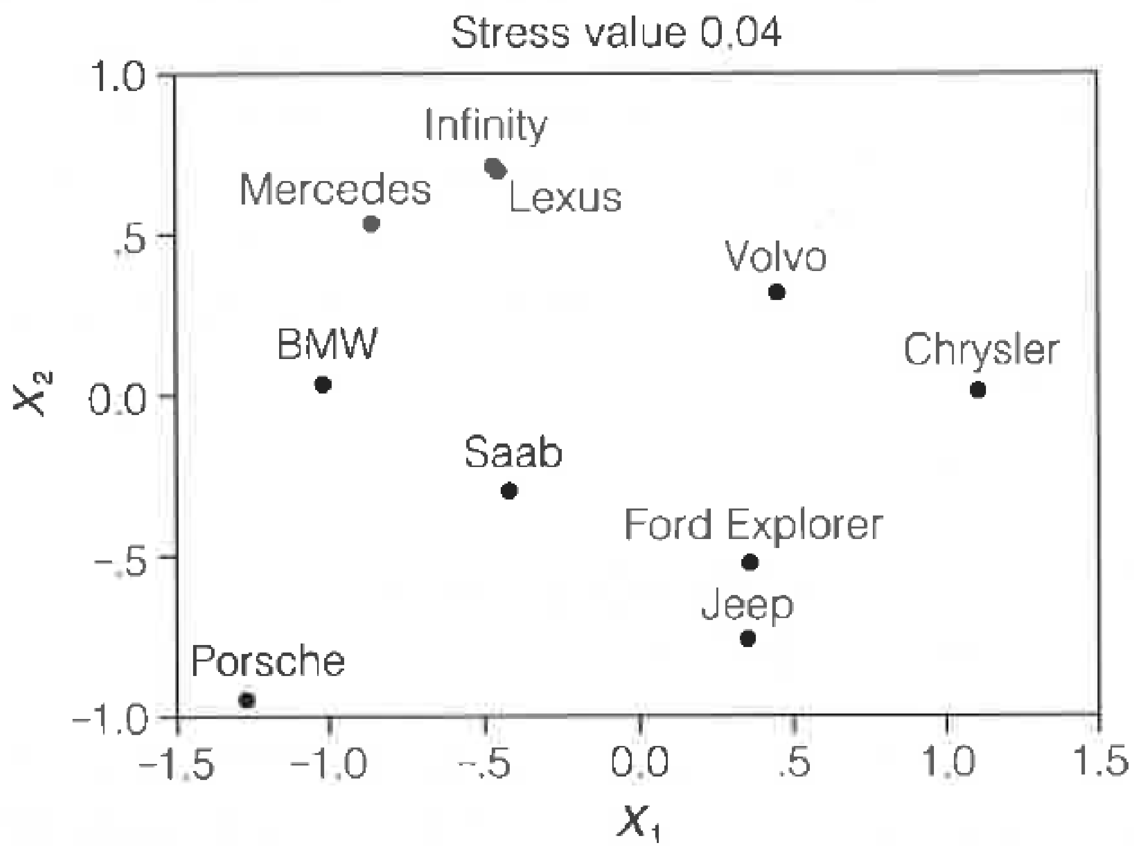
Reflected & Rotated

Stress = 0.06



Non-metric MDS Example

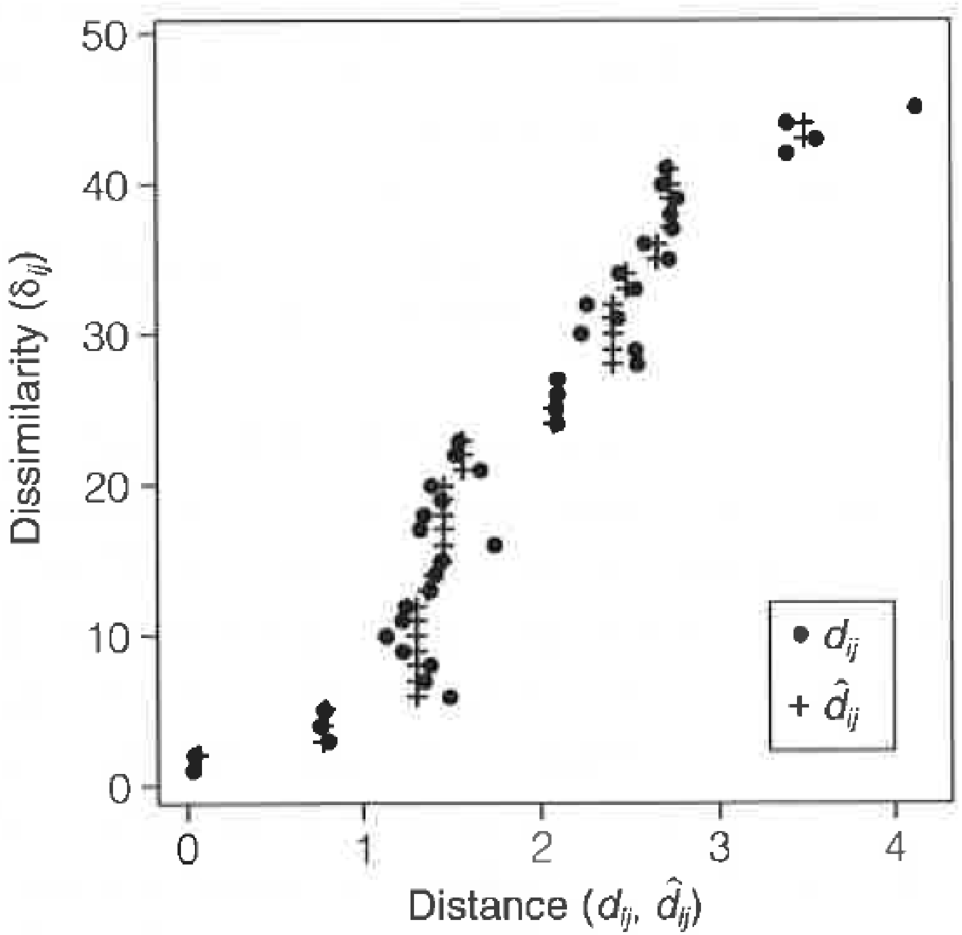
Final Solution
Stress = 0.04



Non-metric MDS Example

FIGURE 7.12
Shepard diagram for best-fitting configuration for car data

Final Solution
Stress = 0.04



Non-metric MDS Example

Reflected & Rotated

Final Solution
Stress = 0.04