Big Data Analysis
(MA60306)

Bibhas Adhikari

Spring 2022-23, IIT Kharagpur

Lecture 12
February 3, 2023

MDS

→ Thus in an MSD processing the input is a similarity matrix and the out put is a low dimensional space, which is usually 2 or 3 dimensional

MDS

→ Thus in an MSD processing the input is a similarity matrix and the out put is a low dimensional space, which is usually 2 or 3 dimensional

Classical MDS In this case, we assume that the data set is

$$\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}, \boldsymbol{x}_j \in \mathbb{R}^d,$$

and the data matrix is

$$X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{d \times n}$$

Note that each column represents a data point here.

MDS

→ Thus in an MSD processing the input is a similarity matrix and the out put is a low dimensional space, which is usually 2 or 3 dimensional

Classical MDS In this case, we assume that the data set is

$$\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}, \boldsymbol{x}_j \in \mathbb{R}^d,$$

and the data matrix is

$$X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{d \times n}$$

Note that each column represents a data point here.
The Euclidean distance matrix is $D = [d_{ij}] = [d_2(\boldsymbol{x}_i, \boldsymbol{x}_j)]$ where

$$d_2(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{j=1}^{n} (x_j - y_j)^2}$$

MDS

We define Euclidean square-distance matrix

$$S = [d_2^2(\boldsymbol{x}_i, \boldsymbol{x}_j)]$$

MDS

We define Euclidean square-distance matrix

$$S = [d_2^2(\boldsymbol{x}_i, \boldsymbol{x}_j)]$$

Then observe that D and S are

- $\triangle$ Symmetric

MDS

We define Euclidean square-distance matrix

$$S = [d_2^2(\boldsymbol{x}_i, \boldsymbol{x}_j)]$$

Then observe that D and S are

- △ Symmetric
- △ invariant under shift and rotation

MDS

We define Euclidean square-distance matrix

$$S = [d_2^2(\boldsymbol{x}_i, \boldsymbol{x}_j)]$$

Then observe that D and S are

- △ Symmetric
- △ invariant under shift and rotation

Euclidean distance metric Any symmetric matrix D is called a Euclidean distance matrix or Euclidean metric if there exists a positive integer k and a set $\mathcal{Z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n\} \subset \mathbb{R}^k$ such that

$$D = [d_2(\boldsymbol{z}_i, \boldsymbol{z}_j)].$$

In that case $\mathcal{Z}$ is called a *configuration* of D.

MDS

Question Can a matrix be detected as a Euclidean metric from its properties without finding the configuration explicitly?

MDS

Question Can a matrix be detected as a Euclidean metric from its properties without finding the configuration explicitly?

Gram (Gramian) matrix of a data set Let $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \subset \mathbb{R}^d$. Then the Gram matrix of $\mathcal{X}$ is defined by

$$G = [g_{ij}] = [\boldsymbol{x}_i^T \boldsymbol{x}_j] = [\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle]$$

MDS

Question Can a matrix be detected as a Euclidean metric from its properties without finding the configuration explicitly?

Gram (Gramian) matrix of a data set Let $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \subset \mathbb{R}^d$. Then the Gram matrix of $\mathcal{X}$ is defined by

$$G = [g_{ij}] = [\boldsymbol{x}_i^T \boldsymbol{x}_j] = [\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle]$$

Then G is a positive semi-definite matrix.

MDS

Question Can a matrix be detected as a Euclidean metric from its properties without finding the configuration explicitly?

Gram (Gramian) matrix of a data set Let $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \subset \mathbb{R}^d$. Then the Gram matrix of $\mathcal{X}$ is defined by

$$G = [g_{ij}] = [\boldsymbol{x}_i^T \boldsymbol{x}_j] = [\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle]$$

Then G is a positive semi-definite matrix. However, every positive semi-definite matrix G has Cholesky decomposition i.e.

$$G = R^T R$$

for some $R = [\boldsymbol{r}_1, \ldots \boldsymbol{r}_n]$

MDS

Question Can a matrix be detected as a Euclidean metric from its properties without finding the configuration explicitly?

Gram (Gramian) matrix of a data set Let $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \subset \mathbb{R}^d$. Then the Gram matrix of $\mathcal{X}$ is defined by

$$G = [g_{ij}] = [\boldsymbol{x}_i^T \boldsymbol{x}_j] = [\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle]$$

Then G is a positive semi-definite matrix. However, every positive semi-definite matrix G has Cholesky decomposition i.e.

$$G = R^T R$$

for some $R = [\boldsymbol{r}_1, \ldots \boldsymbol{r}_n]$

Question What is the conclusion?

MDS

Note that for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$,

$$d_2(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle + \langle \boldsymbol{y}, \boldsymbol{y} \rangle - 2\langle \boldsymbol{x}, \boldsymbol{y} \rangle}$$

i.e.

$$d_{ij} = d_2(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{g_{ii} + g_{jj} - 2g_{ij}}$$

MDS

Note that for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$,

$$d_2(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle + \langle \boldsymbol{y}, \boldsymbol{y} \rangle - 2\langle \boldsymbol{x}, \boldsymbol{y} \rangle}$$

i.e.

$$d_{ij} = d_2(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{g_{ii} + g_{jj} - 2g_{ij}}$$

If the given set of data points $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, $\boldsymbol{x}_j \in \mathbb{R}^d$ lie on a k dimensional affine subspace (hyperplane) $H \subset \mathbb{R}^d$ then the center of

$$\overline{\boldsymbol{x}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i$$

lies in H, and $S = H - \overline{\boldsymbol{x}} = \{x - \overline{\boldsymbol{x}} : x \in H\} \subset \mathbb{R}^d$ is a d-dimensional subspace parallel to H

MDS

Set $\widehat{\mathcal{X}} = \{\widehat{\boldsymbol{x}}_1, \ldots, \widehat{\boldsymbol{x}}_n\}$, $\widehat{\boldsymbol{x}}_i = \boldsymbol{x}_i - \overline{\boldsymbol{x}}$ is called the *centered data set* and the corresponding data matrix $\widehat{X}$ is called the centered data matrix.

MDS

Set $\widehat{\mathcal{X}} = \{\widehat{\boldsymbol{x}}_1, \ldots, \widehat{\boldsymbol{x}}_n\}$, $\widehat{\boldsymbol{x}}_i = \boldsymbol{x}_i - \overline{\boldsymbol{x}}$ is called the *centered data set* and the corresponding data matrix $\widehat{X}$ is called the centered data matrix.
The Gram matrix corresponding to $\widehat{\mathcal{X}}$ given by

$$G^c = [\langle \widehat{\boldsymbol{x}}^T, \widehat{\boldsymbol{x}}_j \rangle] = \widehat{X}^T \widehat{X}$$

is called the *centering Gram matrix* of $\mathcal{X}$.

MDS

Set $\widehat{\mathcal{X}} = \{\widehat{\boldsymbol{x}}_1, \ldots, \widehat{\boldsymbol{x}}_n\}$, $\widehat{\boldsymbol{x}}_i = \boldsymbol{x}_i - \overline{\boldsymbol{x}}$ is called the *centered data set* and the corresponding data matrix $\widehat{X}$ is called the centered data matrix.
The Gram matrix corresponding to $\widehat{\mathcal{X}}$ given by

$$G^c = [\langle \widehat{\boldsymbol{x}}^T, \widehat{\boldsymbol{x}}_j \rangle] = \widehat{X}^T \widehat{X}$$

is called the *centering Gram matrix* of $\mathcal{X}$. If $G^c = [g_{ij}^c]$ then

$$d_{ij} = \sqrt{g_{ij}^c + g_{jj}^c - 2g_{ij}^c}$$

MDS

Set $\widehat{\mathcal{X}} = \{\widehat{\boldsymbol{x}}_1, \ldots, \widehat{\boldsymbol{x}}_n\}$, $\widehat{\boldsymbol{x}}_i = \boldsymbol{x}_i - \overline{\boldsymbol{x}}$ is called the *centered data set* and the corresponding data matrix $\widehat{X}$ is called the centered data matrix.
The Gram matrix corresponding to $\widehat{\mathcal{X}}$ given by

$$G^c = [\langle \widehat{\boldsymbol{x}}^T, \widehat{\boldsymbol{x}}_j \rangle] = \widehat{X}^T \widehat{X}$$

is called the *centering Gram matrix* of $\mathcal{X}$. If $G^c = [g_{ij}^c]$ then

$$d_{ij} = \sqrt{g_{ij}^c + g_{jj}^c - 2g_{ij}^c}$$

(revisiting) Centering/centralizing matrix Let $\mathbf{1} = [1, 1, \ldots, 1]^T \in R^n$. Let $E = \mathbf{1}\mathbf{1}^T$. Then $C_n = I_n - \frac{1}{n}E$ Then

MDS

Then

$\rightarrow C_n^2 = C_n$

MDS

Then

$\rightarrow C_n^2 = C_n$

$\rightarrow \mathbf{1}^T C_n = C_n \mathbf{1} = 0$

MDS

Then

- $\rightarrow$ $C_n^2 = C_n$
- $\rightarrow$ $\mathbf{1}^T C_n = C_n \mathbf{1} = 0$
- $\rightarrow$ A data set $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ is centered if and only if $XC_n = X$, where $X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]$

MDS

Then

- $\rightarrow$ $C_n^2 = C_n$
- $\rightarrow$ $\mathbf{1}^T C_n = C_n \mathbf{1} = 0$
- $\rightarrow$ A data set $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ is centered if and only if $XC_n = X$, where $X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]$
- $\rightarrow$ A positive semi-definite matrix M is a centering Gram matrix if and only if $C_n M C_n = M$

MDS

Then

- $\rightarrow$ $C_n^2 = C_n$
- $\rightarrow$ $\mathbf{1}^T C_n = C_n \mathbf{1} = 0$
- $\rightarrow$ A data set $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ is centered if and only if $XC_n = X$, where $X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]$
- $\rightarrow$ A positive semi-definite matrix M is a centering Gram matrix if and only if $C_n M C_n = M$

Conclusion Let G denote the Gram matrix of a data matrix X. Then the data matrix corresponding to the centered data set of X is XC_n, and the centering Gram matrix of X is $G^c = C_n G C_n$

Note that the data points are columns of X

MDS

Theorem Let $\mathcal{X}$ be a data set. Then

$$G^c = -\frac{1}{2}S^c$$

Proof Note that $\sum_{i=1}^{n} g_{ij}^c = 0$.

MDS

Theorem Let $\mathcal{X}$ be a data set. Then

$$G^c = -\frac{1}{2}S^c$$

Proof Note that $\sum_{i=1}^{n} g_{ij}^c = 0$. Then $d_{ij} = \sqrt{g_{ij}^c + g_{jj}^c - 2g_{ij}^c}$ implies

$$\sum_{i=1}^{n} d_{ij}^2 = ng_{jj}^c + \sum_{i=1}^{n} g_{ii}^c \text{ and } \sum_{j=1}^{n} d_{ij}^2 = ng_{ii}^c + \sum_{j=1}^{n} g_{jj}^c.$$

MDS

Theorem Let $\mathcal{X}$ be a data set. Then

$$G^c = -\frac{1}{2}S^c$$

Proof Note that $\sum_{i=1}^n g_{ij}^c = 0$. Then $d_{ij} = \sqrt{g_{ij}^c + g_{jj}^c - 2g_{ij}^c}$ implies

$$\sum_{i=1}^n d_{ij}^2 = ng_{jj}^c + \sum_{i=1}^n g_{ii}^c \text{ and } \sum_{j=1}^n d_{ij}^2 = ng_{ii}^c + \sum_{j=1}^n g_{jj}^c.$$

Thus

$$\begin{aligned} [S^c]_{ij} &= D_{ij}^2 - \frac{1}{n}\left(\sum_{i=1}^n d_{ij}^2 + \sum_{j=1}^n d_{ij}^2 - \frac{1}{n}\sum_{i,j=1}^n d_{ij}^2\right) \\ &= d_{ij}^2 - g_{ii}^c - g_{jj}^c = -2g_{ij}^c \end{aligned}$$

MDS

Theorem Let $\mathcal{X}$ be a data set. Then

$$G^c = -\frac{1}{2}S^c$$

Proof Note that $\sum_{i=1}^n g_{ij}^c = 0$. Then $d_{ij} = \sqrt{g_{ij}^c + g_{jj}^c - 2g_{ij}^c}$ implies

$$\sum_{i=1}^n d_{ij}^2 = ng_{jj}^c + \sum_{i=1}^n g_{ii}^c \text{ and } \sum_{j=1}^n d_{ij}^2 = ng_{ii}^c + \sum_{j=1}^n g_{jj}^c.$$

Thus

$$\begin{aligned} [S^c]_{ij} &= D_{ij}^2 - \frac{1}{n}\left(\sum_{i=1}^n d_{ij}^2 + \sum_{j=1}^n d_{ij}^2 - \frac{1}{n}\sum_{i,j=1}^n d_{ij}^2\right) \\ &= d_{ij}^2 - g_{ii}^c - g_{jj}^c = -2g_{ij}^c \end{aligned}$$

Question What does this equality mean when the data points are normalized?

MDS

The we have the following consequence: A matrix A is a Euclidean square-distance matrix if and only if $-\frac{1}{2}A^c$ is a centering positive semi-definite matrix.

MDS

The we have the following consequence: A matrix A is a Euclidean square-distance matrix if and only if $-\frac{1}{2}A^c$ is a centering positive semi-definite matrix.

Classical multidimensional scaling method Let $D = [d_{ij}]$ be a given distance matrix for a set of n objects. The we want to find a configuration $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \subset \mathbb{R}^d$ such that a certain distance matrix associated with $\mathcal{X}$ is as close as possible to D, i.e.

$$d_X(\boldsymbol{x}_i, \boldsymbol{x}_j) \approx d_{ij}$$

for $1 \leq i, j \leq n$.

MDS

Lemma Suppose $D = [d_{ij}]$ is an $n \times n$ Euclidean metric and $S = [d_{ij}^2]$ is the corresponding square-distance matrix. Let $G^c = -\frac{1}{2}S^c$. If the rank of G^c is k then there is an k-dimensional centered vector set $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \subset \mathbb{R}^k$ such that

$$d_2(\boldsymbol{x}_i, \boldsymbol{x}_j) = d_{ij}, \ 1 \leq i, j \leq n.$$

MDS

Lemma Suppose $D = [d_{ij}]$ is an $n \times n$ Euclidean metric and $S = [d_{ij}^2]$ is the corresponding square-distance matrix. Let $G^c = -\frac{1}{2}S^c$. If the rank of G^c is k then there is an k-dimensional centered vector set $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \subset \mathbb{R}^k$ such that

$$d_2(\boldsymbol{x}_i, \boldsymbol{x}_j) = d_{ij},\ 1 \leq i, j \leq n.$$

Proof By above G^c is a centering gram matrix. If rank of G^c is k then $G^c = X^T X$ for some matrix $X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{k \times n}$, which has the desired property.

MDS

Lemma Suppose $D = [d_{ij}]$ is an $n \times n$ Euclidean metric and $S = [d_{ij}^2]$ is the corresponding square-distance matrix. Let $G^c = -\frac{1}{2}S^c$. If the rank of G^c is k then there is an k-dimensional centered vector set $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \subset \mathbb{R}^k$ such that

$$d_2(\boldsymbol{x}_i, \boldsymbol{x}_j) = d_{ij}, \ 1 \leq i, j \leq n.$$

Proof By above G^c is a centering gram matrix. If rank of G^c is k then $G^c = X^T X$ for some matrix $X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{k \times n}$, which has the desired property.

Let us call k as the configuration of D and $\mathcal{X}$ as the exact configuration of D. However, note that k can be large to meet the goal that we have low-dimensional configuration, say $d \ll k$.

MDS

Question Can we take help of the PCA?

MDS

Question Can we take help of the PCA?

Let Y denote a random desired matrix. Then consider the loss function

$$\mathcal{L}(\mathcal{Y}) = \sum_{i,j=1}^{n} \left(d_{ij}^2 - d_2^2(\boldsymbol{y}_i, \boldsymbol{y}_j)\right)$$

where $\mathcal{Y}$ is obtained by orthogonal projection from $\mathbb{R}^k$ to a d-dimensional subspace of $\mathbb{R}^k$ and $\mathcal{X}$ is the exact configuration of D.

MDS

Lemma Let $\mathcal{Z} \subset \mathbb{R}^k$ be a given data with Euclidean square-distance matrix $S_Z = [s_{ij}]$, where $s_{ij} = d_2^2(\boldsymbol{z}_i, \boldsymbol{z}_j)$, and let G_Z^c be its centering Gram matrix. Then

$$\operatorname{tr}(G_Z^c) = \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{n} s_{ij}$$

Proof Homework

MDS

Lemma Let $\mathcal{Z} \subset \mathbb{R}^k$ be a given data with Euclidean square-distance matrix $S_Z = [s_{ij}]$, where $s_{ij} = d_2^2(\boldsymbol{z}_i, \boldsymbol{z}_j)$, and let G_Z^c be its centering Gram matrix. Then

$$\text{tr}(G_Z^c) = \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{n} s_{ij}$$

Proof Homework

Lemma Let $D_Z = [d_2(\boldsymbol{z}_i, \boldsymbol{z}_j)]$ and $\widehat{Z} = [\widehat{z}_1, \ldots, \widehat{z}_n]$, the centered data matrix corresponding to $\mathcal{Z}$. Then

$$\|\widehat{Z}\|_F = \frac{1}{\sqrt{2n}} \|D_Z\|_F$$

Proof Homework

MDS

Theorem[1] Let $\mathcal{X} \subset \mathbb{R}^k$ be the configuration of D such that $\mathcal{X}$ is centered and the SVD of X be given by

$$X = U\Sigma_k V^T,$$

where $\Sigma_k = \text{diag}(\sigma_1, \ldots, \sigma_k)$, $U = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_k]$. For a given $d \ll k$, let $U_d = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_d]$ and $Y = U_d^T X$. Then Y is a solution of the above optimization problem with

$$\mathcal{L}(\mathcal{Y}) = \sum_{i=d+1}^{k} \sigma_i^2.$$

Proof Homework

[1] Chapter 6, J. Wang, Geometric Structure of High-Dimensional Data and Dimensionality Reduction, Springer, 2012

MDS

The classical MDS algoritm:

Step 1 Let D be the given distance matrix. Set $G^c = -\frac{1}{2}S^c$

MDS

The classical MDS algoritm:

Step 1 Let D be the given distance matrix. Set $G^c = -\frac{1}{2}S^c$
Step 2 Suppose rank of G^c is k. Compute the spectral decomposition of G^c as $G^c = U\Lambda U^T$, where $U = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k]$, and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_k)$ with $\lambda_1 \geq \lambda_2 \leq \ldots \leq \lambda_k$.

MDS

The classical MDS algoritm:

Step 1 Let D be the given distance matrix. Set $G^c = -\frac{1}{2}S^c$
Step 2 Suppose rank of G^c is k. Compute the spectral decomposition of G^c as $G^c = U\Lambda U^T$, where $U = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k]$, and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_k)$ with $\lambda_1 \geq \lambda_2 \leq \ldots \leq \lambda_k$.
Step 3 Set $U_d = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k]$ and $\Sigma_d = \text{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_d})$. Then the configuration is $Y = \Sigma_d U_d^T$

MDS

```
import numpy as np
from numpy.linalg import eig
D = np.array([[0,4,3,7,8],[4,0,1,6,7],[3,1,0,5,7],[7,6,5,0,1], [8 ,7 ,7 ,1 ,0]])
D2 = np.square(D)
C = np.eye(5) - 0.2*np.ones(5)
M = -0.5* C @ D2 @ C
l,V = eig(M)
s = np.real(np.power(l,0.5))
V2 = V[:,[0,1]]
s2 = np.diag(s[0:2])
Q = V2 @ s2
import matplotlib.pyplot as plt
plt.plot(Q[:,0],Q[:,1],'ro')
plt.show()
```