* Fundamentals of BigData Analytics / 24. Nov. 2017

4.2 Multi-dimensional Scaling

$O_1, \dots, O_n$

$\delta_{ij} = \delta_{ji} \geq 0, \quad \forall (i,j) \quad ; \quad \delta_{ii} = 0$

$\Delta = (\delta_{ij})_{ij}$ dissimilarity matrix

$$\mathcal{U}_n = \left\{ \Delta = (\delta_{ij})_{1 \leq i,j \leq n} \mid \delta_{ij} = \delta_{ji} \geq 0, \ \delta_{ii} = 0 \ \forall i,j = 1,\dots,n \right\}$$

$x_1, \dots, x_n \in \mathbb{R}^K \qquad \|x_i - x_j\| \sim \delta_{ij}$

$X = [x_1 \cdots x_n]^T \in \mathbb{R}^{n \times K} \qquad d_{ij}(X) = \|x_i - x_j\|$

$$\Delta \longleftrightarrow D(X) = \left(\|x_i - x_j\| \right)_{1 \leq i,j \leq n}$$

$$\downarrow$$

$$\Delta^{(q)} = (\delta_{ij}^q)_{ij} \qquad D(X)^{(q)} = \left(\|x_i - x_j\|^q \right)_{ij}$$

Our Goal: $\min_{X \in \mathbb{R}^{n \times K}} \| \Delta^{(q)} - D(X)^{(q)} \|$

$\|\cdot\|$ matrix norm

4.2.1. Characterization of Euclidean Distance Matrix

$D^{(n)}(X)$ $\quad X = [x_1 \dots x_n]^T \in \mathbb{R}^{n \times r}$ $\quad \Delta^{(n)}$

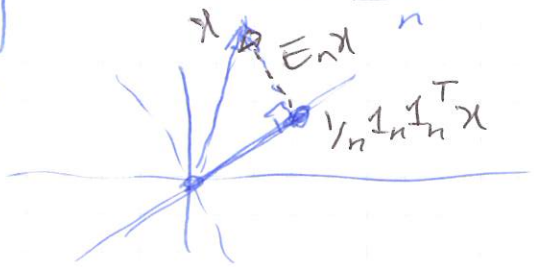
$$E_n = I_n - \frac{1}{n} 1_n 1_n^T \qquad 1_n = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

Centering matrix

$$\frac{1}{n} 1_n \underline{1_n^T x} = \frac{1}{n} 1_n \left(\sum_{i=1}^{n} x_i \right) = \begin{bmatrix} \bar{x} \\ \vdots \\ \bar{x} \end{bmatrix}$$

$$E_n x = x - \begin{bmatrix} \bar{x} \\ \vdots \\ \bar{x} \end{bmatrix} \qquad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\frac{1}{n} 1_n 1_n^T$$



Theorem 4.3. The dissimilarity matrix $\Delta \in \mathcal{U}_n$ has an Euclidean embedding in $\mathbb{R}^K$ if and only if $\frac{-1}{2} E_n \Delta^{(2)} E_n$ is non-negative definite and

$$rk\left(\frac{-1}{2} E_n \Delta^{(2)} E_n\right) \leq K .$$

The least K which allows for an embedding is called dimensionality of Δ.

Proof: $X = [x_1 \dots x_n]^T \in \mathbb{R}^{n \times K}$

$$-\frac{1}{2} D^{(2)}(X) = XX^T - 1_n \hat{x}^T - \hat{x} 1_n^T \quad \text{(exercise)}$$

$$\hat{x} = \frac{1}{2} [x_1^T x_1 \dots x_n^T x_n]^T.$$

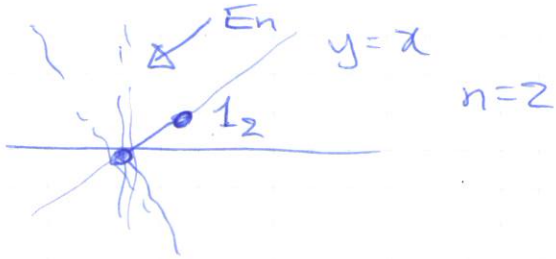
Hint: $\|x_i - x_j\|^2 = (x_i - x_j)^T (x_i - x_j)$

$$= x_i^T x_i + x_j^T x_j - 2 x_i^T x_j$$

$$-\frac{1}{2} E_n D^{(2)}(X) E_n = E_n X X^T E_n$$

$E_n 1_n = 0$

$1_n^T E_n = 0$



first: $-\frac{1}{2} E_n D^{(2)}(X) E_n = E_n X (E_n X)^T \geq 0$

second: $rk(-\frac{1}{2} E_n D^{(2)}(X) E_n) = rk(E_n X X^T E_n) \leq K.$

$n \times n \quad n \times K \quad K \times n \quad n \times n$

$n \geq n \geq K \geq n \geq n$

$K \leq n$

if there is X for diss.. matrix Δ: $\Delta^{(2)} = D^{(2)}(X)$

$$\Rightarrow -\frac{1}{2} E_n \Delta^{(2)} E_n = -\frac{1}{2} E_n D^{(2)}(X) E_n$$

① n.n.d.

② $rk(-\frac{1}{2} E_n D^{(2)} E_n) \leq K$

$-\frac{1}{2} E_n \Delta^{(2)} E_n \succeq 0 \qquad rk\left(-\frac{1}{2} E_n \Delta^{(2)} E_n\right) \leq K$

$X \succeq 0$

$$-\frac{1}{2} E_n \Delta^{(2)} E_n = E_n X X^T E_n$$

$$(E_n X)(E_n X)^T$$

There exists $n \times K$ matrix X such that

$$-\frac{1}{2} E_n \Delta^{(2)} E_n = X X^T \quad X^T E_n = X^T \text{ (Proof)}$$

Then $X = [x_1 \cdots x_n]^T$ is an appropriate configuration: $-\frac{1}{2} E_n D^{(2)}(X) E_n = E_n X X^T E_n = X X^T$

$$= -\frac{1}{2} E_n \Delta^{(2)} E_n .$$

□

Step 1) find $-\frac{1}{2} E_n \Delta^{(2)} E_n$

Step 2) See if it is n.n.d. and check the rank.

Step 3) if it is non-negative definite Then find X: $-\frac{1}{2} E_n \Delta^{(2)} E_n = X X^T$.

Hint: $A \succeq 0 \Rightarrow A = V \Lambda V^T = \underbrace{(V \Lambda^{1/2})}_{X} (\Lambda^{1/2} V^T)$

4.2.2. The Best Euclidean fit to a given dissimilarity matrix.

$\Delta \in \mathcal{U}_n \quad \alpha \longrightarrow D(X)$

notation $\lambda^+ = \max(\lambda, 0)$

Theorem 4.4. Let $\Delta \in \mathcal{U}_n$ be a dissimilarity matrix and $-\frac{1}{2} E_n \Delta^{(2)} E_n$ has the spectral decomposition : $-\frac{1}{2} E_n \Delta^{(2)} E_n = V \operatorname{diag}(\lambda_1 \cdots \lambda_n) V^T$ with $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_n$ and V an orthogonal matrix.

Then the optimization problem

$$\min_{X \in \mathbb{R}^{n \times k}} \| E_n (\Delta^{(2)} - D^{(2)}(X)) E_n \|$$

has a solution given by

$$X^* = \left[\sqrt{\lambda_1^+} v_1, \cdots, \sqrt{\lambda_k^+} v_k \right] \in \mathbb{R}^{n \times k}$$

Proof. Note that a solution to

$$\min_{\substack{A \succeq 0 \\ rk(A) \leq k}} \| -\frac{1}{2} E_n \Delta^{(2)} E_n - A \| \qquad \text{(Thm. 2.6)}$$

is given by

$$A^* = V \operatorname{diag}(\lambda_1^+, \cdots, \lambda_k^+, 0, \cdots 0) V^T.$$

$$-\frac{1}{2} E_n D^{(2)}(X^*) E_n = E_n X^* X^{*T} E_n$$

$$= E_n [v_1 \cdots v_k] \operatorname{diag}(\lambda_1^+, \cdots \lambda_k^+) \times [v_1 \cdots v_k]^T E_n$$

$$= V \operatorname{diag}(\lambda_1^+, \cdots \lambda_k^+, 0 \cdots 0) V^T = A^*$$

$$\Rightarrow \quad X^* \longrightarrow D^{(2)}(X^*) \longrightarrow -\frac{1}{2} E_n D^{(2)}(X^*) E_n = A^*$$

$$X^* \longrightarrow \underset{\text{minimizer}}{\| -\frac{1}{2} E_n (\Delta^{(2)} - D^{(2)}(X)) E_n \|}$$

D

Step 1) $-\frac{1}{2} E_n \Delta^{(2)} E_n$

Step 2) Spectral decomposition

Step 3) $[\sqrt{\lambda_1^+} v_1, \cdots \sqrt{\lambda_k^+} v_k]$

4.2.3 non-linear Dimensionality Reduction

- Manifold learning
- Geodesic

Isomap : complete isometric feature mapping (2000)

$x_1, \dots x_n \in \mathbb{R}^p$

1) Construct a neighborhood graph: find a weighted Graph $G(V, E, W)$

$v_i = x_i$. Two vertices v_i and v_j are connected only if $\|x_i - x_j\| \leq \varepsilon$.

(another way is to connect each point to its K-nearest neighbors)

2) Compute the shortest paths for each pair (v_i, v_j) (Dijkstra's algorithm) .

The geodesic distance $\delta_{ij} = \delta(v_i, v_j)$ can be taken as number of hops/lines from v_i to v_j or sum of $\|x_\ell - x_k\|$ on the shortest path.

3) Construct d-dimensional embedding - Apply MDS on the basis of geodesic distances $\Delta = (\delta_{ij})_{1 \leq i,j \leq n}$

1) Very large distance may distort local neighbor hoods

2) Computational Complexity

3) Not robust to noisie : choice of "ε"

4.3) Diffusion maps.

Weights of $G(V, E, W)$

weight functions or Kernel $w_{ij} = K(x_i, x_j)$

The selected Kernel should satisfy three properties

- Symmetry $K(x_i, x_j) = K(x_j, x_i)$
- non-negative $K(x_i, x_j) \geq 0$
- Locality : there is a scale parameter ε

such that if $\|x_i - x_j\| \ll \varepsilon$ then $K(x_i, x_j) \to 1$

if $\|x_i - x_j\| \gg \varepsilon$ then $K(x_i, x_j) \to 0$

Gaussian Kernel:

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\varepsilon^2}\right)$$

Caveat: Do not confuse this "Kernel" with Kernels of support vector machines (later!)