Multi dimensional scaling

input: $n\times n$ distance matrix $d \in \mathbb{R}_+^{n\times n}$ for convenience let $D_{ij} := d_{ij}^2$

goal: find embedding $y_1,\dots,y_n \in \mathbb{R}^k$ for some (small) k
such that $\|y_i - y_j\|^2 \approx D_{ij}$

explicit optimization ansatz: fix k, choose initial $y_1,\dots y_n \in \mathbb{R}^k$
try to minimize mismatch

$$E(y_1,\dots,y_n) := \frac{1}{2}\sum_{ij}\left(\|y_i - y_j\|^2 - D_{ij}\right)^2$$

problem: E (and similar functionals) are non-convex, difficult to minimize.

classical "spectral" MDS: if such an embedding exists, it can be recovered by matrix diagonalization.

for now assume: D induced by Euclidean embedding
$D_{ij} = \|x_i - x_j\|^2$ for some $x_1,\dots,x_n \in \mathbb{R}^k$ for some k

can we recover coordinates $(x_i)_i$ from D?
- can only be possible up to rotations and translations

Gram matrix: let $A \in \mathbb{R}^{n\times n}$ with $A_{ij} := 1$, $C := id - \frac{1}{n}A$

- set $G := -\frac{1}{2}\,CDC = -\frac{1}{2}\left(id - \frac{1}{n}A\right)D\left(id - \frac{1}{n}A\right)$
  $= -\frac{1}{2}\left(D - \frac{1}{n}AD - \frac{1}{n}DA + \frac{1}{n^2}ADA\right)$
- we will now use following notation:
  - $x_i^2 = \|x_i\|^2$ • $x_i x_j = x_i^T x_j$
  - $\langle x\rangle = \frac{1}{n}\sum_i x_i$ $\quad \langle x^2\rangle = \frac{1}{n}\sum_i x_i^2$
- by translation invariance can choose: $\langle x\rangle = 0$ (data is centered)
- observation: $\frac{1}{n}\sum_i x_i x_j = \langle x\rangle \cdot x_j = 0$

- with these conventions find:

$$\left(\frac{1}{n} AD\right)_{ij} = \frac{1}{n} \sum_k D_{kj} = \frac{1}{n} \sum_k \left(x_k^2 - 2x_k x_j + x_j^2 \right)$$

$$= \langle x^2 \rangle - \underbrace{2 \langle x \rangle x_j}_{=0} + x_j^2$$

$$\left(\frac{1}{n} DA\right)_{ij} = \dots = \langle x^2 \rangle + x_i^2$$

$$\left(\frac{1}{n^2} ADA\right)_{ij} = \frac{1}{n^2} \sum_{kl} D_{kl} = \frac{1}{n^2} \sum_{kl} \left(x_k^2 - 2x_k x_l + x_l^2 \right)$$

$$= 2 \langle x^2 \rangle$$

$$\Rightarrow \quad G_{ij} = -\frac{1}{2}\left(D_{ij}^2 - \langle x^2 \rangle - x_i^2 - \langle x^2 \rangle - x_j^2 + 2 \langle x^2 \rangle \right)$$

$$= x_i^T x_j$$

- now let $X \in \mathbb{R}^{n \times k}$ ← # of dimensions

  # of samples

  then: $G_{ij} = \sum_{a=1}^{k} X_{ia} X_{ja} = X X^T \in \mathbb{R}^{n \times n}$

  covariance matrix: $\mathrm{cov}_{ab} = \frac{1}{n} \sum_i X_{ia} X_{ib}$

  so $\mathrm{cov} = \frac{1}{n} X^T X \in \mathbb{R}^{k \times k}$

- let v, λ be eigvec / eigval of cov.

$$\frac{1}{n} X^T X v = \mathrm{cov}\, v = \lambda \cdot v \quad \Rightarrow \quad X X^T X v = n \lambda X v$$

⇒ then Xv is eigenvector of G with eigenvalue $n \cdot \lambda$

- so what we do is closely related to PCA.
- let now $(z_a, \lambda_a)_a$ be all eigenvectors / value of G

  claim: set $y_{ia} := z_{ai} \sqrt{\lambda_a}$ gives a valid embedding of X

- build inner product matrix from y:

$$y_i^T y_j = \sum_a y_{ia} y_{ja} = \sum_a z_{ai} \lambda_a z_{aj} = G_{ij}$$

if we sort eigenvalues in decreasing order, can think of entries of y_i as increasingly accurate approximations.

- if D was NOT generated by Euclidean embedding (and no such embedding exists), then this cannot work.
- this manifests itself as negative eigenvalues in G
- in this case we can take largest positive eigenvalues in G to obtain approximate embeddings.