12. Metric Scaling


**Given:** $n \times n$ matrix $\Delta$ of dissimilarities between $n$ objects ($\Delta^T = \Delta; \delta_{ij} \geq 0; \delta_{ii} = 0$).

**Goal:** Find points $y_1, \ldots, y_n$ such that euclidean inter-point distances $d(x_i, x_j)$ closely reflect dissimilarities.

**Note:** The points $y_1, \ldots, y_n$ can only be determined up to shift and rotation.

**Note:** Will have to define what we mean by “closely reflect”: Need to define function $\Phi(\Delta, D)$ measuring difference between dissimilaries and distances (stress of the configuration).

**Note:** Alternatively we might be given $n \times n$ matrix $\Theta$ of simliarities ($\Theta^T = \Theta; \theta_{ij} \leq \theta_{ii}$).

Can convert similarities into dissimilarities by setting

$$\delta_{ij}^2 = \theta_{ii} + \theta_{jj} - 2\theta_{ij}.$$
We already know:

Suppose we are given $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^m$.
Set $\delta_{ij} = d(\mathbf{x}_i, \mathbf{x}_j)$ euclidean interpoint distance.
Let $y_1, \ldots, y_n$ be obtained by projecting $\mathbf{x}_1, \ldots, \mathbf{x}_n$ onto a $k$-D subspace. Set $d_{ij} = d(y_i, y_j)$.
Then the subspace minimizing

$$\Phi(\Delta, D) = \sum_{ij} (\delta_{ij}^2 - d_{ij}^2)$$

is spanned by the $k$ largest principal components (eigen-vectors of $\Sigma$).

What if:

- We are not given $\mathbf{x}_1, \ldots, \mathbf{x}_n$, but only $\Delta$?
- We do not know whether $\Delta$ is euclidean (a euclidean interpoint distance matrix for a set of points in some euclidean space.)
Fact: Can easily

- Check whether $\Delta$ is euclidean;
- If yes, find out in which dimension, and
- Construct $x_1, \ldots, x_n$ with interpoint distance matrix $\Delta$.

How? Suppose $x_1, \ldots, x_n \in R^m$ for some $m$ and $\delta_{ij} = d(x_i, x_j) = \|x_i - x_j\| \Rightarrow$

$$\delta_{ij}^2 = \|x_i\|^2 + \|x_j\|^2 - 2 \langle x_i, x_j \rangle.$$

Look at matrix with elements $\delta_{ij}^2$.
Sweep out row means $\Rightarrow (i, j)$-th element is

$$\|x_j\|^2 - \frac{1}{n} \sum_j \|x_j\|^2 - 2 \langle x_i, x_j - \bar{x} \rangle.$$

Sweep out column means $\Rightarrow (i, j)$-th element is

$$-2 \langle x_i - \bar{x}, x_j - \bar{x} \rangle.$$
So: Denote by $A$ the $n \times n$ matrix with elements $a_{ij} = -\frac{1}{2} \delta_{ij}^2$. Define
$$H = I - \frac{1}{n} \mathbf{1} \mathbf{1}^T.$$  
If $\Delta$ is interpoint distance matrix of $n$ points in $m$-D space then $HAH$ is centered inner product matrix $\rightarrow HAH$ is positive semidefinite with rank $\leq m$.

Compute eigen-decomposition of $B = HAH$:
$$B = HAH = U \Lambda U^T.$$  
Suppose $\lambda_1, \ldots, \lambda_q > 0$ and $\lambda_{q+1}, \ldots, \lambda_n = 0$.
Let $X$ be the $n \times q$ matrix with $i$-th columns $X_i = \sqrt{\lambda_i} U_i$.
Then the rows of $X$, regarded as points in $q$-D space, have interpoint distance matrix $\Delta$.

Why:
$$\|x_i - x_j\|^2 = b_{ii} + b_{jj} - 2b_{ij}$$
$$b_{ij} = a_{ij} - a_i - a_j + a.. \Rightarrow$$
$$\|x_i - x_j\|^2 = a_{ii} + a_{jj} - 2a_{ij} \quad \text{because } A^T = A$$
$$= -2a_{ij} \quad \text{because } a_{ii} = a_{jj} = 0$$
$$= \delta_{ij}^2$$