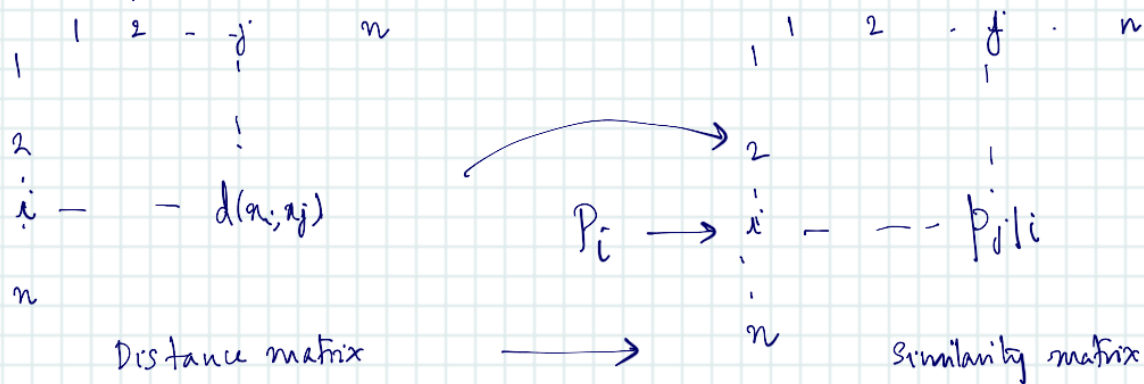


- Recap
- SNE
- Recall: Create a map or an embedding of high dim. data points in a low dim. space such that the structure in the high dim. space is retained as much as possible in the low dim. space
- $X = \{x_1, x_2, \dots, x_n\}$ is the data set consisting of high dim. data points x_i
- $Y = \{y_1, y_2, \dots, y_n\}$ is the map consisting of low dim. map points y_i
- Construct pair wise distances in the high dim. space



$$\underline{p_{ij}} = \underline{p_{j|i}} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / 2\sigma_i^2)}$$

Assume that the distances map to probabilities under a Gaussian prior

- Similarly, we define a distance matrix and a similarity matrix on the map points y_i

$$\underline{q_{ij}} = \underline{q_{j|i}} = \frac{\exp(-\|y_i - y_j\|^2)}{\sum_{k \neq i} \exp(-\|y_i - y_k\|^2)}$$

- Claim: If the structure in the data points is retained in the map points, this is reflected in the similarity of the conditional distributions.

• Let's define the cost function $C = \sum_{i=1}^n KL(P_i \| Q_i)$ — (1)

* Reading exercise: We know the $KL(p||q)$ is asymmetric. What is its impact on the map?

• Find $\gamma = \{\gamma_1, \gamma_2, \dots, \gamma_n\}$ such that C is minimized.

• Iterative gradient descent based solution.

$$\gamma^{(t)} = \gamma^{(t-1)} - \eta \nabla_{\gamma^{(t-1)}} C + \alpha(t) (\gamma^{(t-1)} - \gamma^{(t-2)}) \quad - (2)$$

$$\frac{\partial C}{\partial \gamma_i} = 2 \sum_j (p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j}) (\gamma_i - \gamma_j)$$