① $p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / 2\sigma^2)}$, $p_{i|i} = 0$, $p_{ij} = \frac{p_{i|j} + p_{j|i}}{2}$

• $e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}} = \sqrt[2\sigma^2]{e^{\|x_i - x_j\|^2}}$ ⇒ consider an increase in σ to be a change of scale in the probabilities, the neighbour graph plots should be identical but with a darker shade of gray when $\sigma = 5$

⇒ high similarity should result in darker groups ⇒ we should see 3 blockes in the neighbour graph plots

⇒ only two of them have darker blocks along the diagonal (b) & f))

⇒ since $\sigma = 5$ should be darker ⇒ { b) → $\sigma = 5$; f) → $\sigma = 2$

② if $k < D$, we usually have less degrees of freedom to capture all of the relevant features in the input data. So when applying this bottleneck in the reconstruction networks, the most prominent features are learned and the rest are discarded. Since we lose some amount of information, although we are able to reconstruct most of the information, the data which was discarded can't be reconstructed, therefore we incur some reconstruction loss in our model.

• if there exists correlated data in an input we can model the full information even with the bottleneck layer. Consider two parts of the input x_1 and x_2 where we have $x_2 = \lambda x_1$ with some arbitrary $\lambda \in \mathbb{R}$: the autoencoder, even with less neurons will be able to reconstruct the original information since two data points can be reconstructed with a learnt feature and different weights.
So, as long as the number of correlated data points $\geq D - k$ the autoencoder should be able to have zero loss

formally: $f(x) = X W_1 W_2$ and $X \in \mathbb{R}^D$, $W_1 \in \mathbb{R}^{D \times k}$, $W_2 \in \mathbb{R}^{k \times D}$
so as long as X can be fully represented in $\mathbb{R}^k$ with zero overlap between points the autoencoder will have zero loss.