Math 690 F2017: Topics in Data Analysis and Computation
Homework 3

Xiuyuan Cheng

This problem set asks to compare (a) Isomap, (b)LLE, (c) Laplacian eigenmap, and (d) t-SNE to synthetic manifold data and handwritten digits data.

1. Generate data by uniformly sampling $n = 1000$ points on a closed curve embedded in $\mathbb{R}^3$, e.g.

$$(\cos(2\pi t), \sin(2\pi t), \cos(8\pi t))^T, \quad 0 \le t \le 1.$$

A typical output is shown in Figure below.

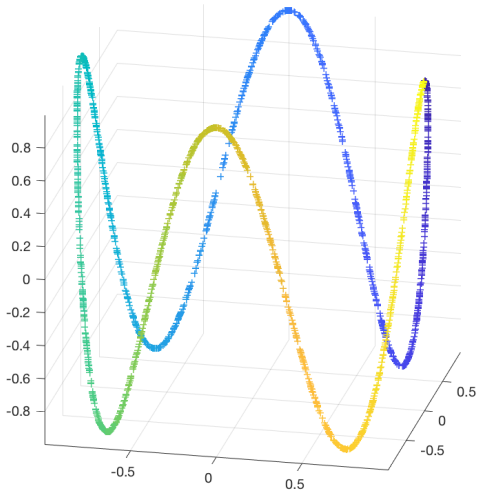

Figure 1: Manifold data, no noise.

(1) Effect of n: what are the embedding like for the (a)-(d) methods? What the effect of having n increased e.g. doubled?

(2) Effect of parameter: in (a)-(c), what is the effect of choosing different k in constructing the kNN-graph? Similarly, what is the effect of choosing different ε in constructing the ε-graph in (a),(c)? For (d), the corresponding parameter of k is the "perplexity".

(3) Effect of noise: what if add Gaussian noise $\mathcal{N}(0, \sigma^2 I_3)$ to the data? What is the influence of the noise level σ?

(4) Effect of ambient dimension: what if the curve is embedded in higher dimension than 3, e.g. in dimension 10 or 30? How does Gaussian noise change the embedding?

(5) Effect of topology: what if changing the curve into a belt (intrinsic dimension 2)? How does the width of the belt affect the embeddings? Question(*): What happens if it is a Mobius belt?

2. Load the handwritten digits for example MNIST and repeat (1)-(3) of Problem 1. To simplify visualization, you can consider fewer number of digits than 10, for example, 5 classes only. Or you can consider 3 classes where 2 of them are "close", for example, '3' and '8', '4' and '9'.