1 EM Algorithm on Manifolds

1.1 Assumptions

Suppose the data are scattered in k different clusters on a d-manifold, and for each cluster i, they are generated from a geodesic Gaussian distribution with the pdf:

$$f_i(x) \propto \frac{1}{\sigma_i^d} e^{-\frac{d_G^2(x,\mu_i)}{d\sigma_i^2}}$$

where μ_i denotes the mean of the distribution, σ_i^2 as the variance and $d_G(x,y)$ denotes the corresponding geodesic distance between two points x and y on the manifold.

1.2 Algorithms

Given N data points setteled on the underlying manifold, and K different clusters are assumed on it. Moreoevr, suppose we have can well represent the geodesic distance between any two arbitrary data points x and x' on the manifold, $d_G(x,x')$, our EM algorithm is designed as such:

Initialize K different centers $\mu_1, \cdots \mu_k$ selected from all N data points

E-Step

for $j = 1, 2, \cdots, n$

1. *assign the cluster label* $y_j = \arg\min_{1 \leq j \leq k} \frac{1}{(\sigma_i^2)^{\frac{d}{2}}} e^{-\frac{d_G^2(x_j,\mu_i)}{d\sigma_i^2}}$

M-Step

for $i = 1, 2, \cdots, k$

1. $\pi_i = \frac{|C_i|}{\sum_{j=1}^{k} |C_j|}$
2. $\mu_i = \arg\min_{x \in C_i} \sum_{x_j \in C_i} d_G^2(x, x_j)$
3. $\sigma_i^2 = \frac{\sum_{x_j \in C_i} d_G^2(x_j,\mu_i)}{|C_i|}$

Algorithm 1: EM for Manifold Clustering

1.3 Ways for Constructing Geodesic Distance

In my implementation of the algorithm, we have constructed the geodesic distance by first constructing a k-nearest neighbour graph or an ϵ-neighbour graph where the corresponding weights of the edges of the graph is their Euclidean distance and then replacing the pairwise geodesic distances by the shortest path.

2 Experimental Results

2.1 Example One

The data points were first generated from four clusters on a 2-dimensional space as such:

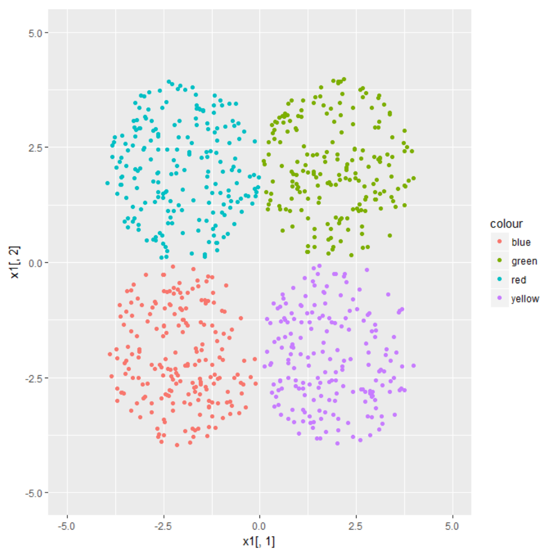

Figure 1: Manifold Data Before Folded

Then all data points on the 2-dimensional space is then folded into a 2-manifolds in a 3-dimensional Euclidean space and adding a Gaussian noise, which looks like:

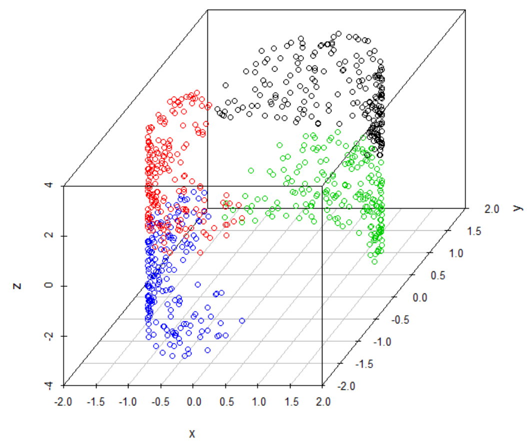

Figure 2: Manifold Data After Folded

Then we conducted both the traditional EM algorithm and the manifold EM algorithm to cluster the manifold data points, and the following results are:

Figure 3: Left: Clustering Result of EM; Right: Clustering Result of Manifold EM