Is spectral clustering mode preserving ?

Stéphane Chrétien*,**, Kavya Jagan **

*ERIC Laboratory and UFR ASSP, University of Lyon 2, 5 avenue Mendes France,
69676 Bron Cedex,
https://sites.google.com/site/stephanegchretien/home
stephane.chretien@univ-lyon2.fr
**National Physical Laboratory
Hampton Road
Teddington TW11 0LW, UK
stephane.chretien,kavya.jagan@npl.co.uk

Clustering is an essential task in machine learning and data science. Standard clustering methods suffer from various drawbacks such as resorting to non-convex optimisation, lack of scalability, need for separation between clusters. We propose a clustering method which involves embedding data using Laplacian Eigenmaps Belkin et Niyogi (2003) and finding the modes in the lower dimensional representation of the data. Laplacian eigenmaps are scalable embedding techniques corresponding to a relaxation of the Maximum Variance Unfolding method Weinberger et Saul (2006). In order to assess the ability of the method to work in cases where clustering overlap, we demonstrate through simulations that Laplacian eigenmaps approximately preserve modes of the high dimensional data in the embedded space. As a consequence of the mode preserving property of Laplacian Eigenmaps, this method is as good as finding modes in the original high dimensional space but with much better computational efficiency. The approach circumvents the curse of dimensionality in the case where the original data lie on a low dimensional submanifold of the possibly high dimensional data space.

In an industrial context, the method offers several advantages :

— the method gives a result in polynomial computational time,
— the result does not depend on the initialisation of the procedure, as opposed to model based estimation procedures. Hence, the result is always the same at each run of the algorithm,
— this method is applicable in cases where the between cluster separation is small,
— the affinity kernel can be defined based on the specific application and hence the method offers flexibility.

Figure 1a shows random Gaussian data in $\mathbb{R}^2$. The stars represent the modes of the clusters used to generate the data. The data is transformed onto $\mathbb{R}^2$ using Laplacian eigenmaps and the modes in the original data are tracked as shown in Figure 1b. It can be observed that the modes in the embedded space lie roughly in the centre of the embedded clusters and that the pairwise distances between data points are maintained on average.

A similar experiment was conducted in which random Gaussian data was generated on a unit sphere and embedded into $\mathbb{R}^2$ space. Figure 2 shows that even in the case where data lie in a higher dimensional manifold, Laplacian eigenmaps preserve the modes of the original