A graph showing different colored dots

Description automatically generatedA graph showing different colored dots

Description automatically generatedMNIST Example – Real Labels (left) and K-means Clustering (right):

**1) Top Right Cluster**

A black silhouette of a city

Description automatically generatedA screen shot of a graph

Description automatically generatedA graph showing different colored dots

Description automatically generatedThe k-means algorithm incorrectly split the top right cluster into two.

The path between the two clusters exhibits minimal complexity and does not pass through any other clusters. There is also little variation in the path weights, which are all relatively small (path weights are scaled relative to the largest path weight in the MST). This means the orange and red points are tightly packed, suggesting they belong to the same cluster. Note the path travels along the outskirts of the cluster because the path is naturally pulled towards the rest of the data points.

**A black silhouette of a city

Description automatically generatedA graph with colored dots

Description automatically generated with medium confidenceA graph showing a graph of colored dots

Description automatically generated with medium confidence2) Bottom Middle Cluster**

The path between the two clusters is complex and passes through another cluster (the orange/red cluster). The low-dimensional embedding and 2D path projection suggest the orange/red clusters is a significant distance away from the turquoise/purple cluster. This, however, isn’t supported by the plot of path weights. The small weights correspond to the portion of the path in the orange/red cluster, while the large weights correspond to the portions of the path in the turquoise cluster (leftmost weights) and the purple cluster (rightmost weights). Notice the segments between clusters are no longer than the segments within the turquoise and purple clusters. This exaggeration of distance is a result of UMAP’s nonlinearity and the inability to accurately represent the global structure of the data in two dimensions.

The mesh of turquoise and purple points in the 2D path projection is not enough to conclude these points belong to the same cluster. Due to the complexity of the path, the PCA projection of the path only contains 56% of the path’s original variance. It’s possible there’s separation between the turquoise and purple points that was reduced by the projection.

If we choose different points from the turquoise and purple clusters, their close relation to each other becomes more obvious.

A graph with dots and lines

Description automatically generatedA graph showing different colored dots

Description automatically generatedThis path is much more direct and suggests the turquoise and purple points belong to the same cluster.

**A black silhouette of a city

Description automatically generated**A graph of a line drawn on a white background

Description automatically generatedA graph with dots and lines

Description automatically generated**3) Turquoise Cluster**

The turquoise cluster was split into two separate sub-clusters. If we highlight both sub-clusters and examine the path between their medoids, we see clear separation in the 2D path projection plot. Since a linear projection can only diminish separation, this proves separation between the two clusters exist in high dimension as well. UMAP was correct to split up the turquoise cluster.

**A) Importance of trying multiple points**

The MST is rather sparse, meaning it doesn’t always depict the simplest path between two points. Since these paths are quite noisy, it is important to test multiple points when studying the relationship between two clusters. The discussion on the turquoise/purple cluster is a great example of this. Even the path between nearby points can be significantly different from the path between the original pair of points. These differences can also be exaggerated by nonlinear dimension reduction methods. In the turquoise/purple cluster discussion, the orange/red cluster is much closer than depicted, so the differences between the two discussed paths were heavily exaggerated in the low-dimensional embedding.

The intra-cluster structure can also be heavily distorted in the low-dimensional embedding. When studying the relationship between two clusters, it’s natural to study the path between their nuclei. However, the center of a cluster in low dimension will not always correspond with the center of the corresponding cluster in high dimension. The brush tool provides a great way to find the low-dimensional point corresponding the medoid of the high-dimensional cluster. Paths between high-dimensional medoids better capture the relationship between clusters.

**B) General Workflow**

Suppose you’re interested in studying the relationship between cluster A and cluster B.

1. Use the brush tool to visualize the path between their medoids. A direct path that does not pass through any other clusters with uniform path weights suggests the two clusters belong to the same sub-population. A complex path that passes through other clusters is not enough evidence to conclude the two clusters belong to separate sub-populations. MST paths can be quite noisy and erratic.
2. Examine the 2D path projection. If inter-cluster separation exists in the projection, then inter-cluster separation exists in high dimension, implying the two clusters belong to separate sub-populations. No separation is not enough evidence to conclude the clusters belong to the same sub-population. The separation may have been diminished by the PCA projection, especially when the projection only retains a fraction of the original variance.
3. The path weights can provide an undistorted depiction of cluster density and global structure. Nonlinear DR techniques like UMAP are known to distort cluster sizes/densities and inter-cluster distances. The high-dimensional path weights provide a more faithful representation of these aspects than the low-dimensional embedding.
4. It is worth checking pairs of points other than the cluster medoids. Nearby points can have drastically different looking paths due to nonlinear DR techniques’ ununiform transformation of space.