CS573 Data Mining
Homework 5

Fangda Li
li1208@purdue.edu

April 27, 2017

1 Exploration

Figure 1: Visualization of digits.

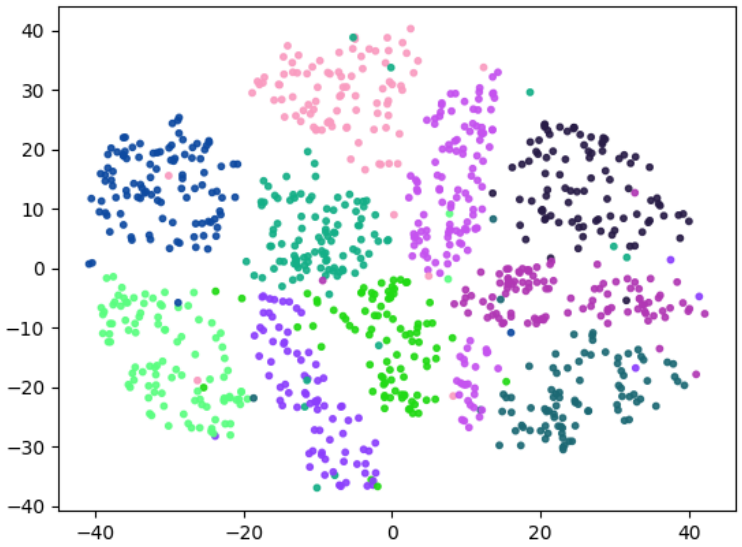
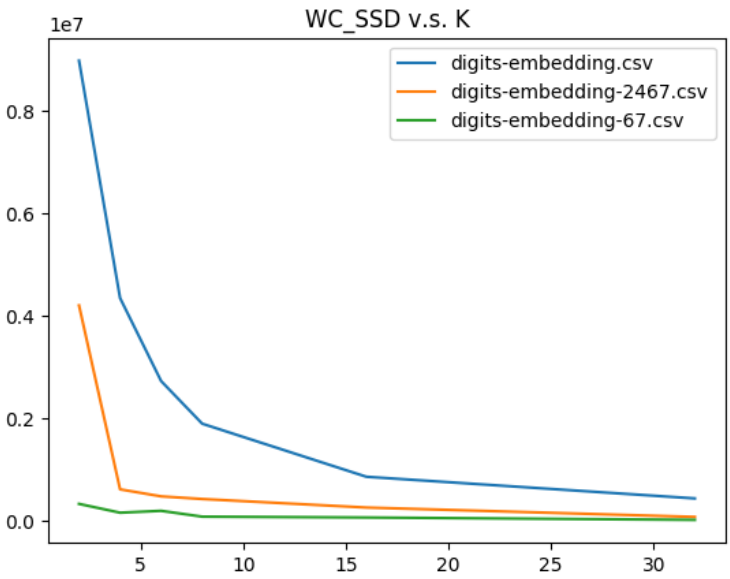


Figure 2: Visualization of 2D tSNE embeddings for 1000 digits.

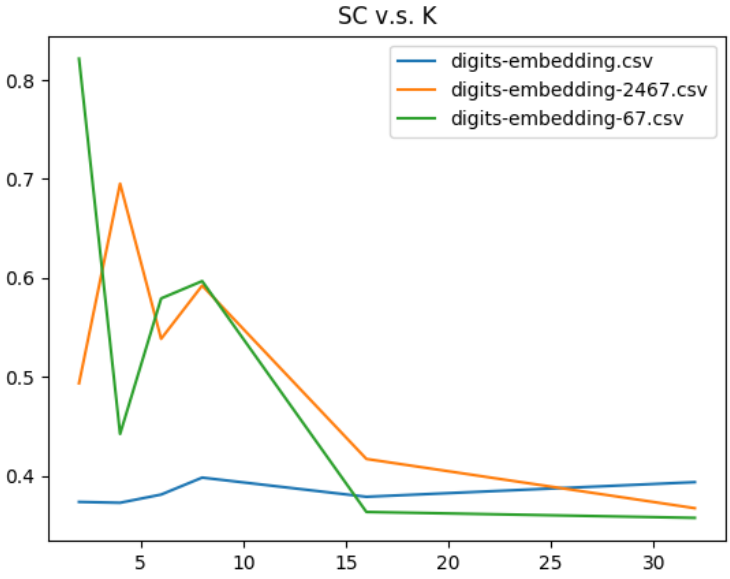
2 Analysis of K-Means

Table 1: NMI for the three versions of data.

	$K = 8$, `embedding`	$K = 4$, `embedding-2467`	$K = 2$, `embedding-67`
NMI	0.347	0.455	0.491

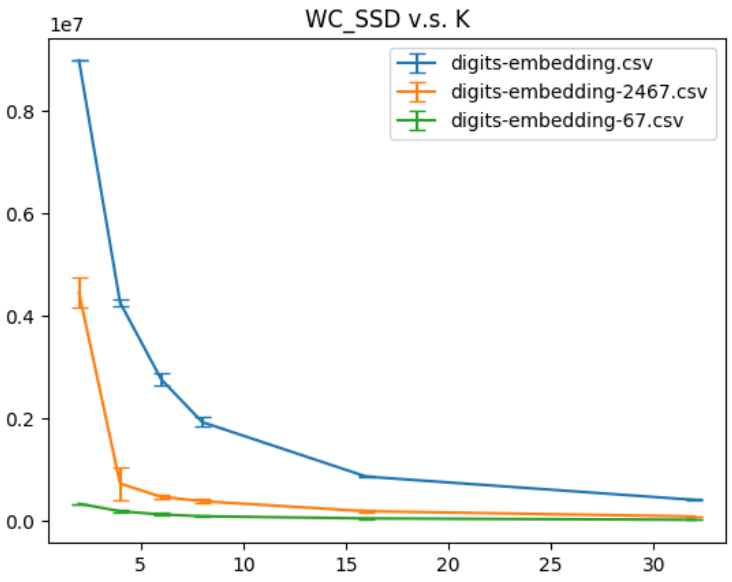


(a) WC SSD v.s. K using k-means.

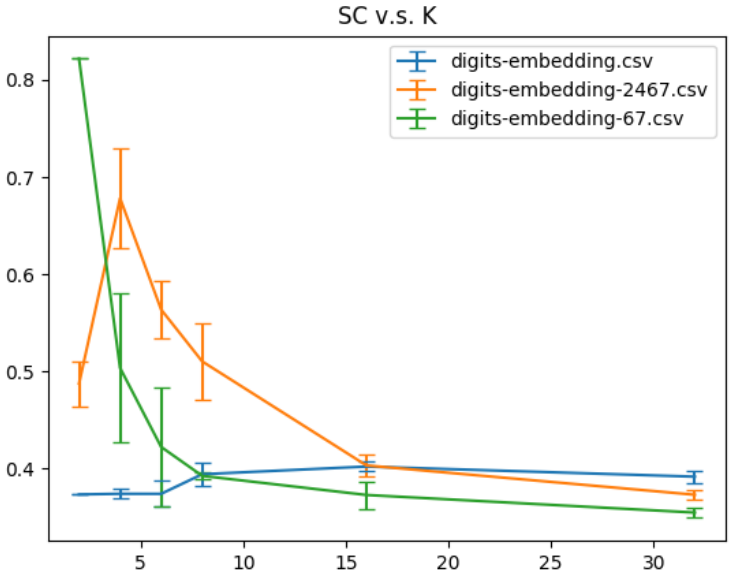


(b) SC v.s. K using k-means.

Figure 3: As K increases, the radii of k-means clusters decreases. Not surprisingly, the Within Cluster Sum of Squared Distance (WC SSD) is a monotonic decreasing function of K. As a result, we have to determine the best K for each dataset by selecting the K with the highest SC score. It is then shown $K = 8, 4, 2$ are chosen for `digits-embedding.csv`, `digits-embedding-2467.csv`, `digits-embedding-67.csv`, respectively. The choices of K align with our intuition that each digit should have its own cluster.

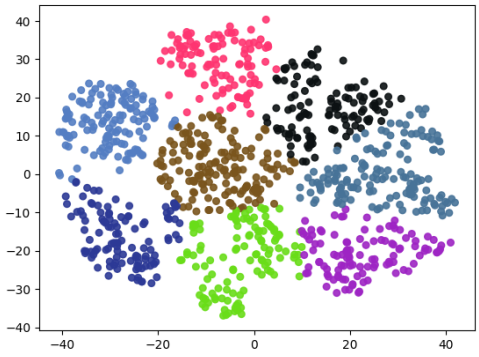


(a) WC SSD v.s. K using k-means over 10 trials.

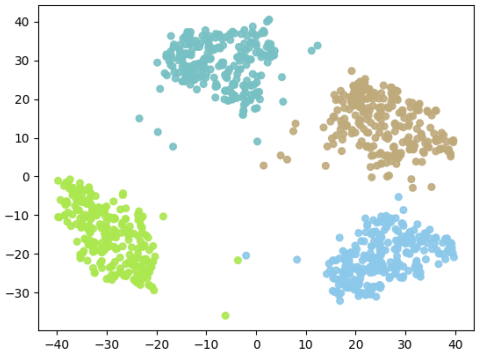


(b) SC v.s. K using k-means over 10 trials.

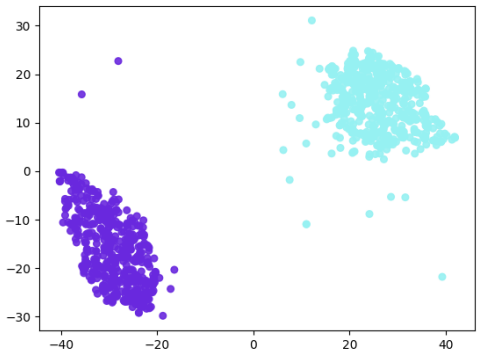
Figure 4: First note that the error bar on each data point represents the standard deviation across the 10 trials. In the WC SSD plot, the variances are not significant. Whereas in the SC plot, we can observe that k-means is more invariant to initial conditions when K is very small (e.g. 2) or large (e.g. 16 and 32), than when K is intermediate (e.g. 4, 6 and 8). We can also recognize that among the three versions of data, `digits-embedding.csv`, the version with the most classes, is most insensitive to the starting conditions.



(a) $K = 8$ for `digits-embedding.csv`



(b) $K = 4$ for `digits-embedding-2467.csv`



(c) $K = 2$ for `digits-embedding-67.csv`

Figure 5: Visualization of clustering results on the three versions of data. We can visually identify that the clusters are well separated on the two smaller datasets. On the contrary, the clusters on `digits-embedding.csv` are all close to each other with debatable boundaries. We can quantitatively validate our observation in Table 3, where the SC scores for the second and third dataset are noticeably higher than the score of the first.

3 Comparison to Hierarchical Clustering

Table 2: Average NMI across 10 trials for the three methods of linkage.

	$K = 32$, single	$K = 16$, complete	$K = 16$, average
NMI	0.370	0.398	0.409

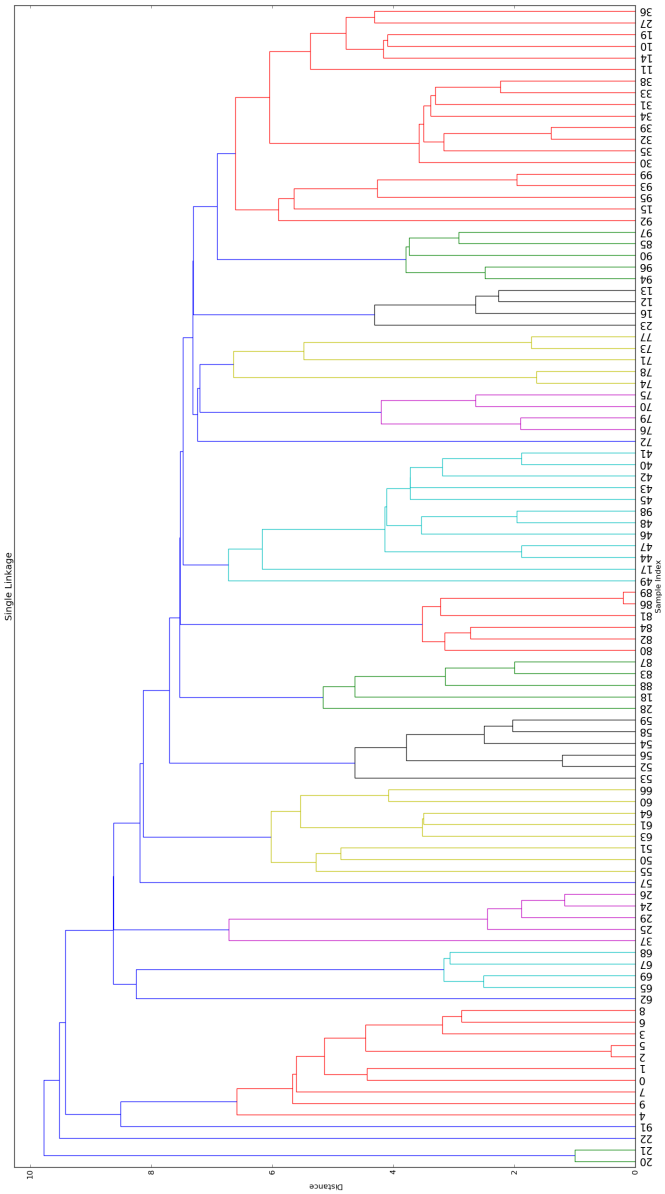


Figure 6: Dendrogram using single linkage.

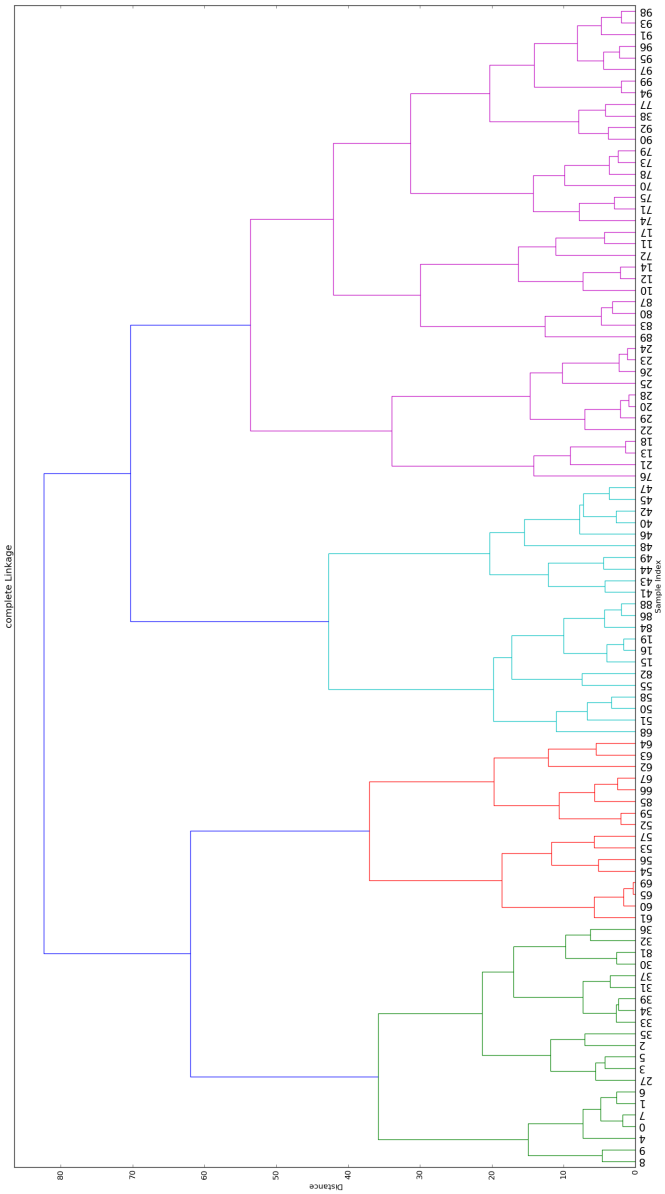


Figure 7: Dendrogram using complete linkage.

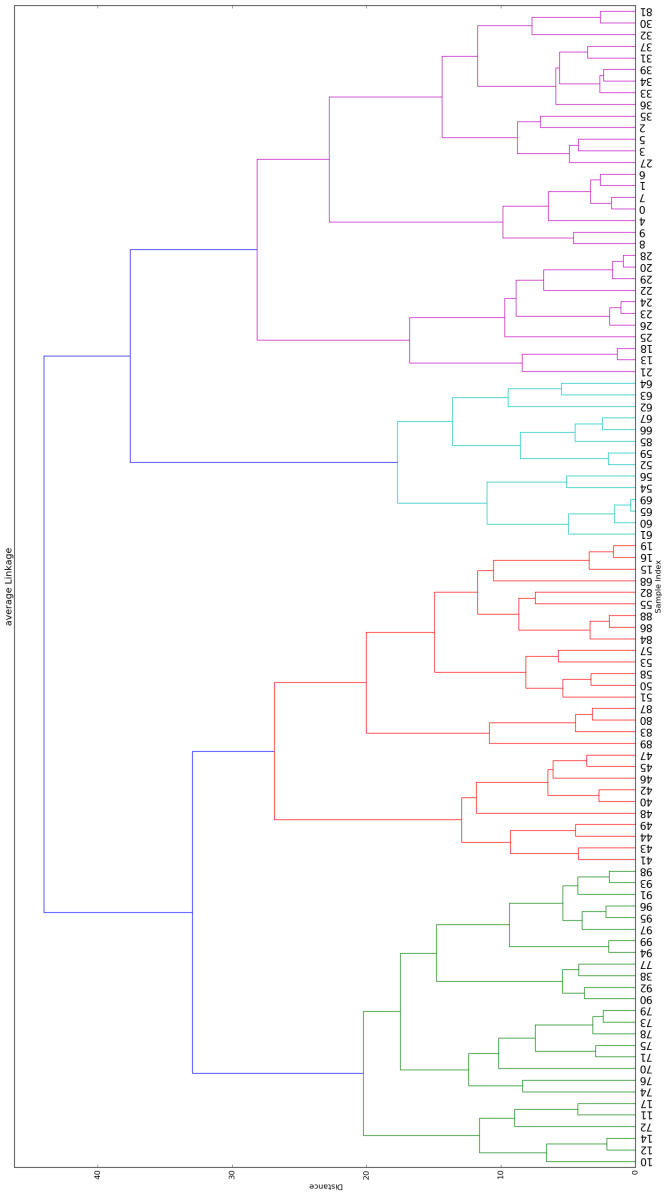
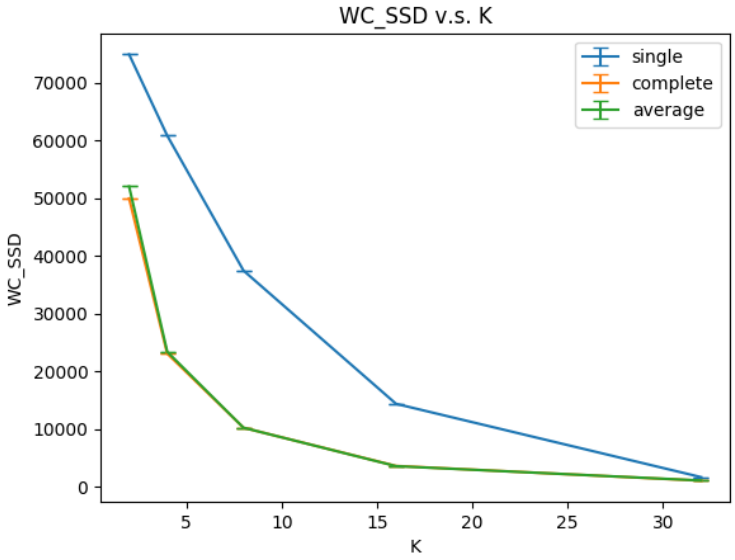
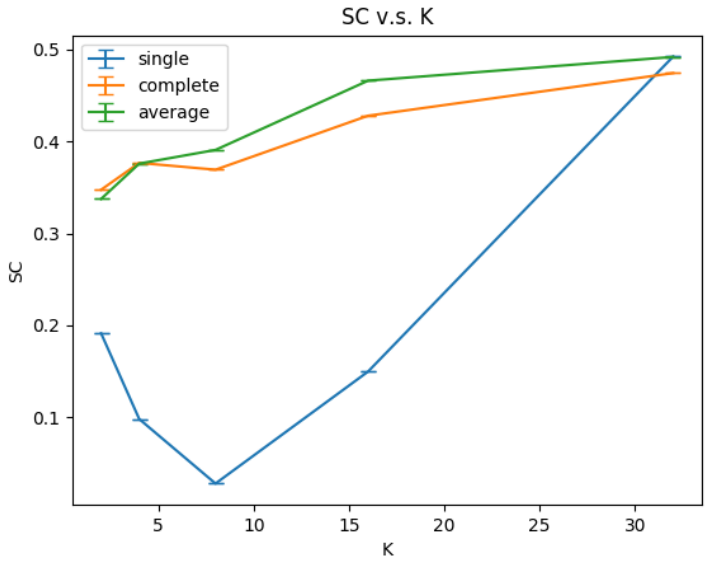


Figure 8: Dendrogram using average linkage.



(a) WC SSD v.s. K using agglomerative clustering.



(b) SC v.s. K using agglomerative clustering.

Figure 9: The candidates for K should be knee points in the WC SSD curve and have a relatively high SC score. Therefore, $K = 32, 16, 16$ have been chosen for single, complete and average linkage, accordingly. The choices for K are different (higher) than those of part B for the full dataset. The NMI scores for each linkage method and its corresponding choice of K are shown in Table 2. Among the three experiments, $K = 16$ with average linkage achieves the highest NMI score. Since we do not run k-means and agglomerative clustering on identical datasets, it is difficult to make a totally fair comparison. However, we can observe that the NMI scores of agglomerative clustering are higher than those of k-means of $K = 8$ on `digits-embedding.csv`.

4 Bonus

Table 3: NMI for the three versions of data.

	$K = 8$, `pca-embedding`	$K = 4$, `pca-embedding-2467`	$K = 2$, `pca-embedding-67`
NMI	0.347	0.319	0.457

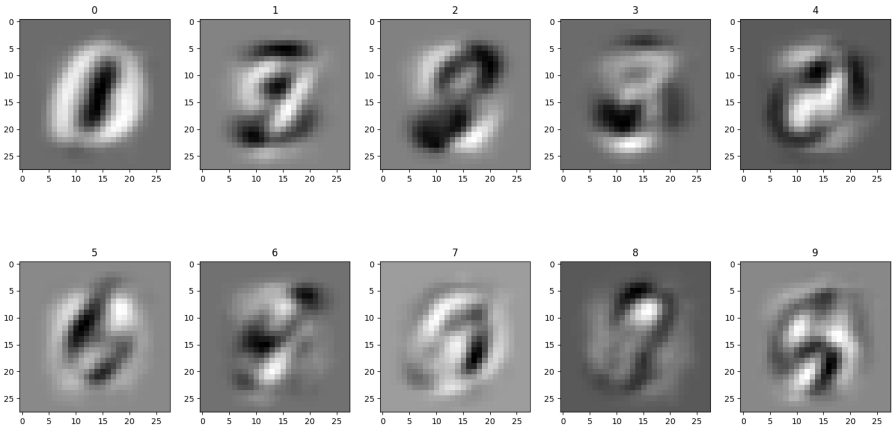


Figure 10: Top 10 axes from PCA (reshaped).

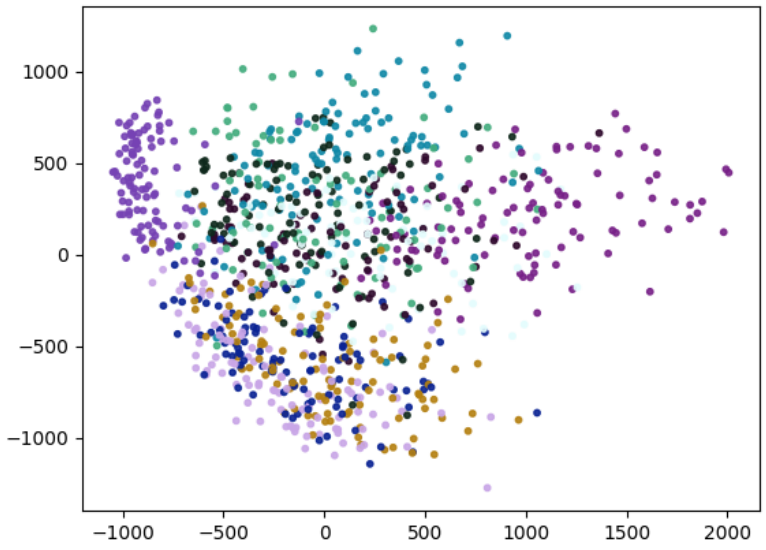
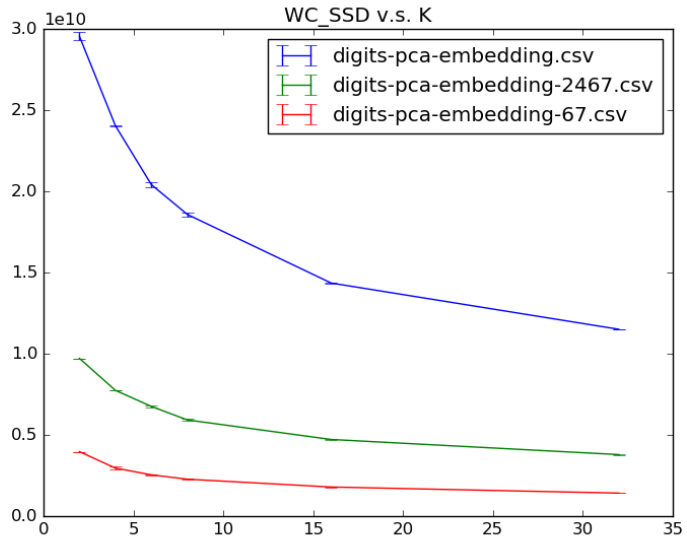
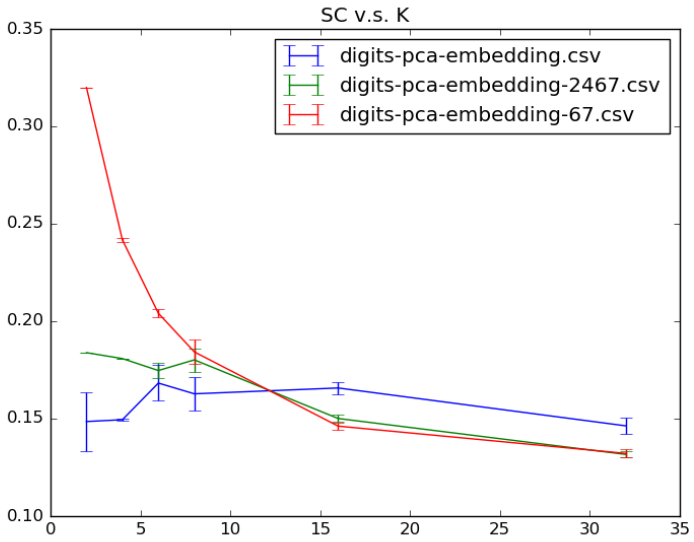


Figure 11: Samples projected onto the first two principle components, where each sample point is colored by its class label. Compared to tSNE embedding, the projections using only the top two principle axes are not discriminative.

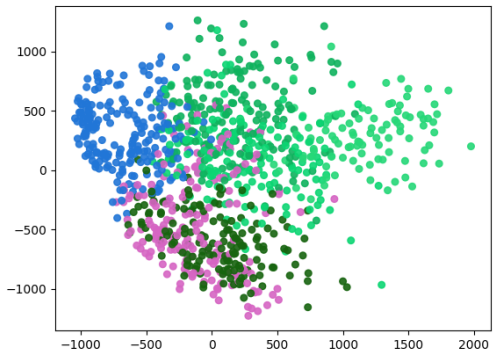


(a) WC SSD v.s. K using k-means on PCA embeddings.

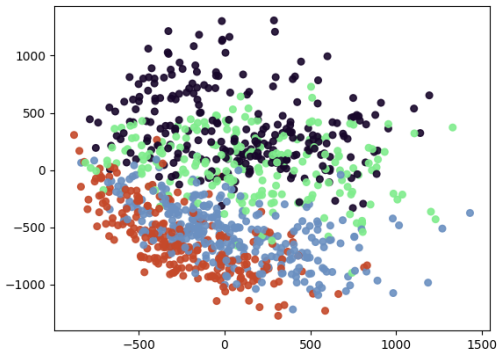


(b) SC v.s. K using k-means on PCA embeddings.

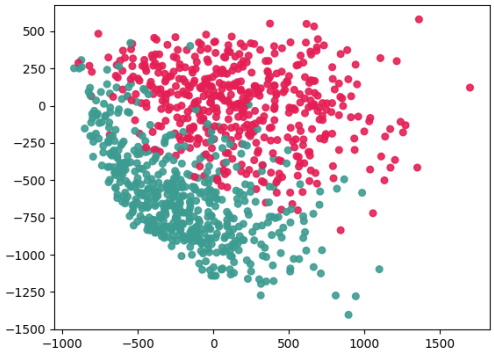
Figure 12: As K increases, the radii of k-means clusters decreases. Not surprisingly, the Within Cluster Sum of Squared Distance (WC SSD) is a monotonic decreasing function of K. As a result, we have to determine the best K for each dataset by selecting the K with a high SC score and/or being one of the knee points in the WC SSD curve. It is then shown $K = 6, 4, 2$ are chosen for `digits-pca-embedding.csv`, `digits-pca-embedding-2467.csv`, `digits-pca-embedding-67.csv`, respectively. The choices of K are slightly different from those of tSNE embedding but they are as intuitive.



(a) $K = 6$ for `pca-embedding.csv`



(b) $K = 4$ for `pca-embedding-2467.csv`



(c) $K = 2$ for `pca-embedding-67.csv`

Figure 13: Visualization of clustering results (projected onto the top two principle axes) on the three versions of data.