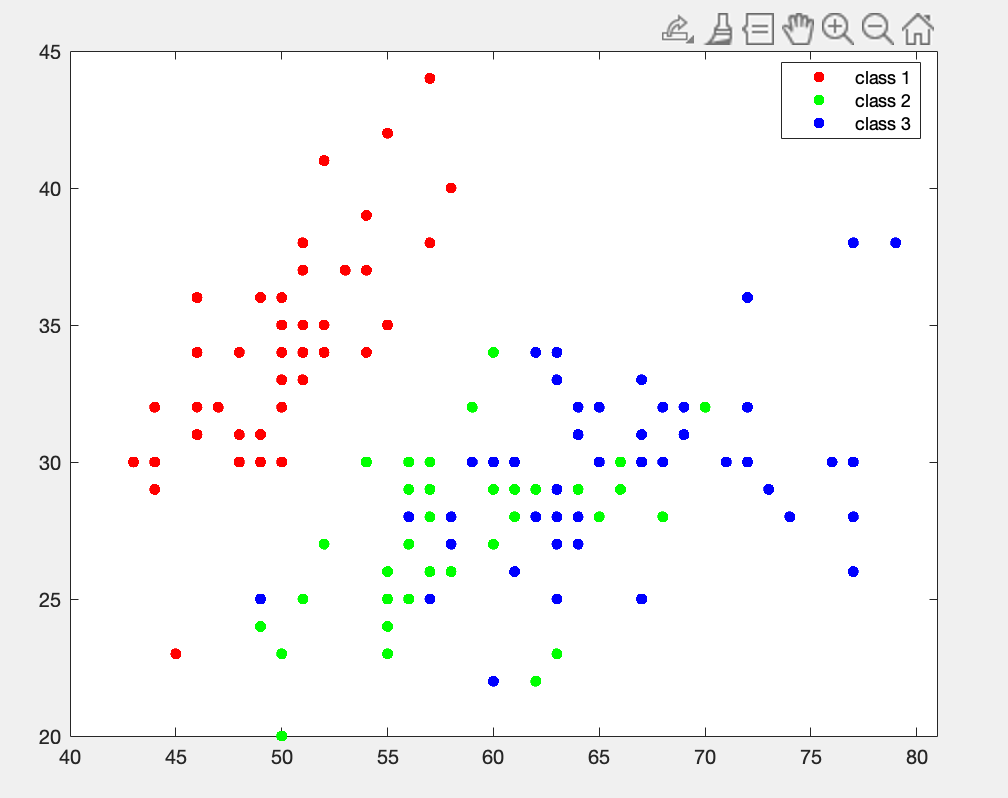
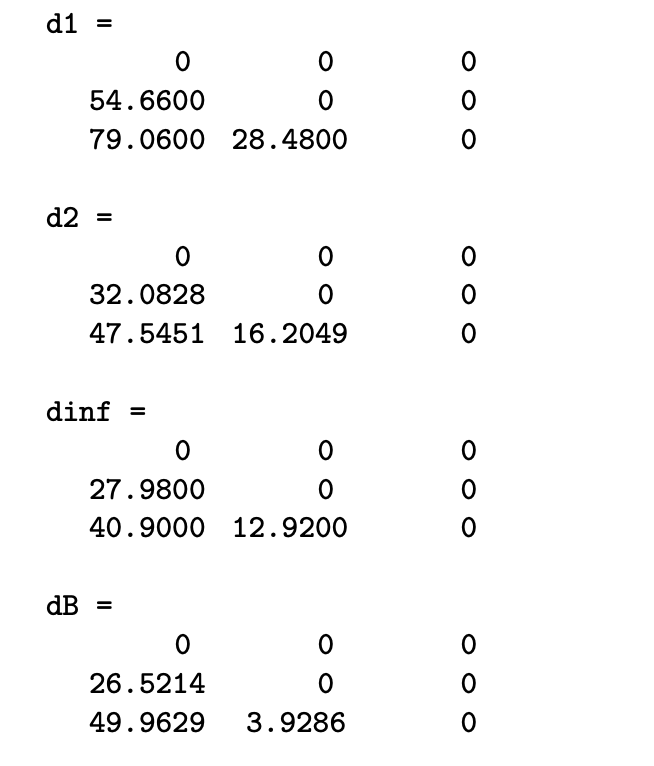
# Report —— Problem Set 1

**Chapter 9**





* The classes are well-separated between class1 and the others as evidenced by the clear separation of the clusters in the scatter plot. On the other hand, class2 and class3 are difficult to separate.
* The gap between class 1 and the others are obvious, which means class 1 is easy to separate linearly with class 2 and 3
* The distances (dp1, dp2, dpinf, dB) can also prove that, the distances between class1 and class 2, class 3 are large enough, however, the distance between class 2 and class 3 is too small, which is difficult to separate.

By observing the distance matrices as follow:  


In all d1, d2, dinf and dB, class2 and class3 have the farthest distance (79.06,47.5451,40.9,49.9629), then is class1 and class2 (54.66, 32.0828, 27.98, 26.5214), and class1 and class3 (28.48,16.2049,12.92,3.9286) have the shortest distance.

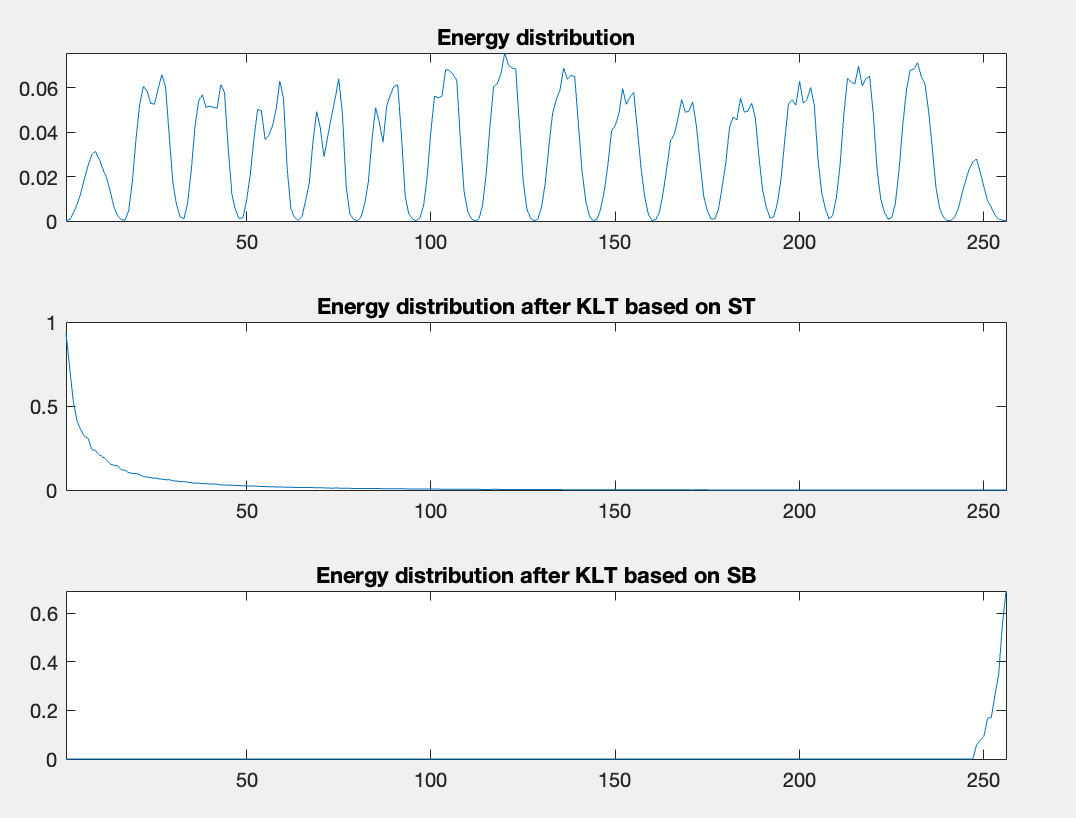
SB=

SW=

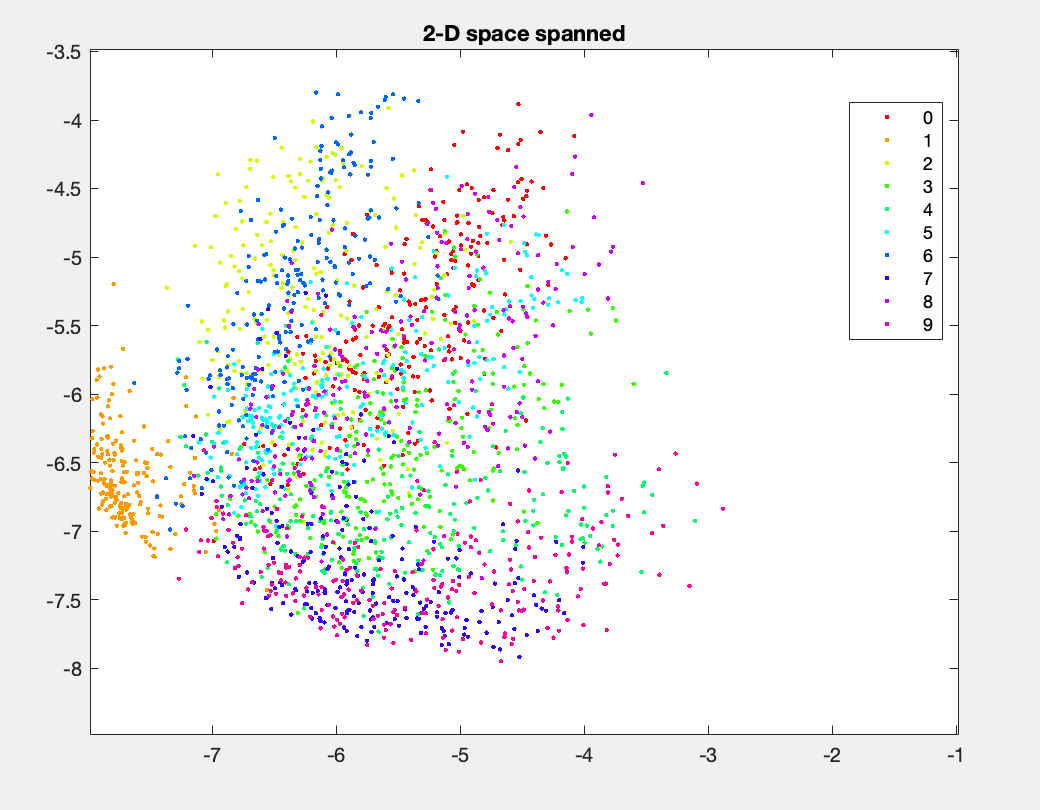
ST=

**Chapter 10**

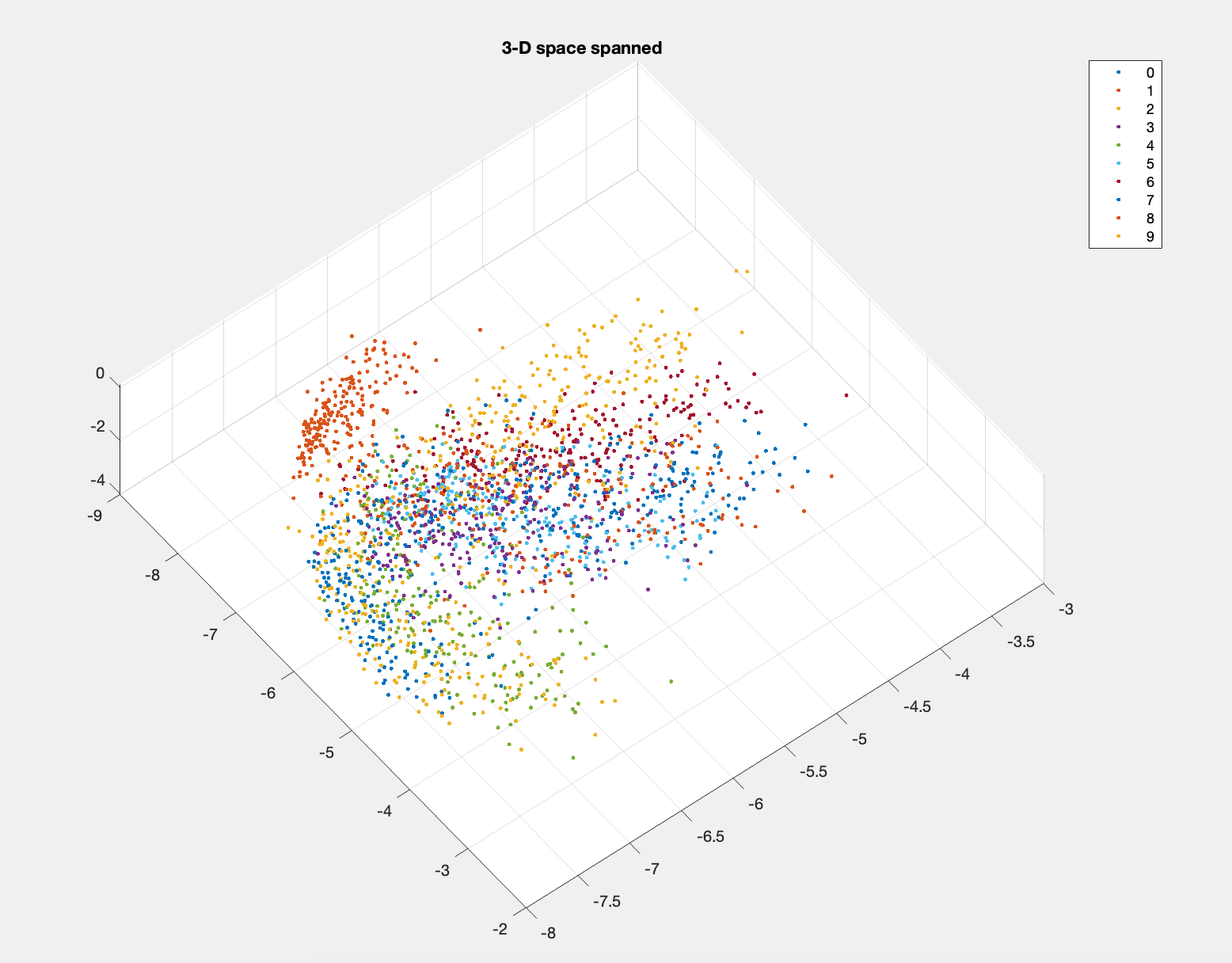
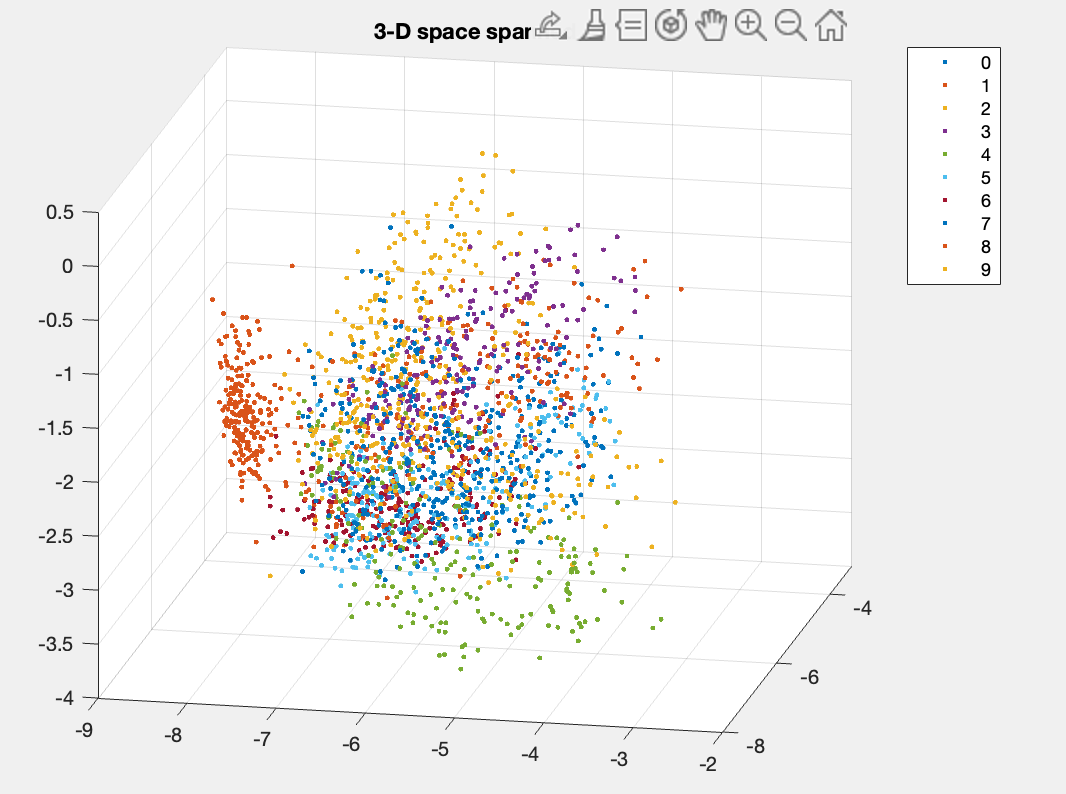
energy distribution



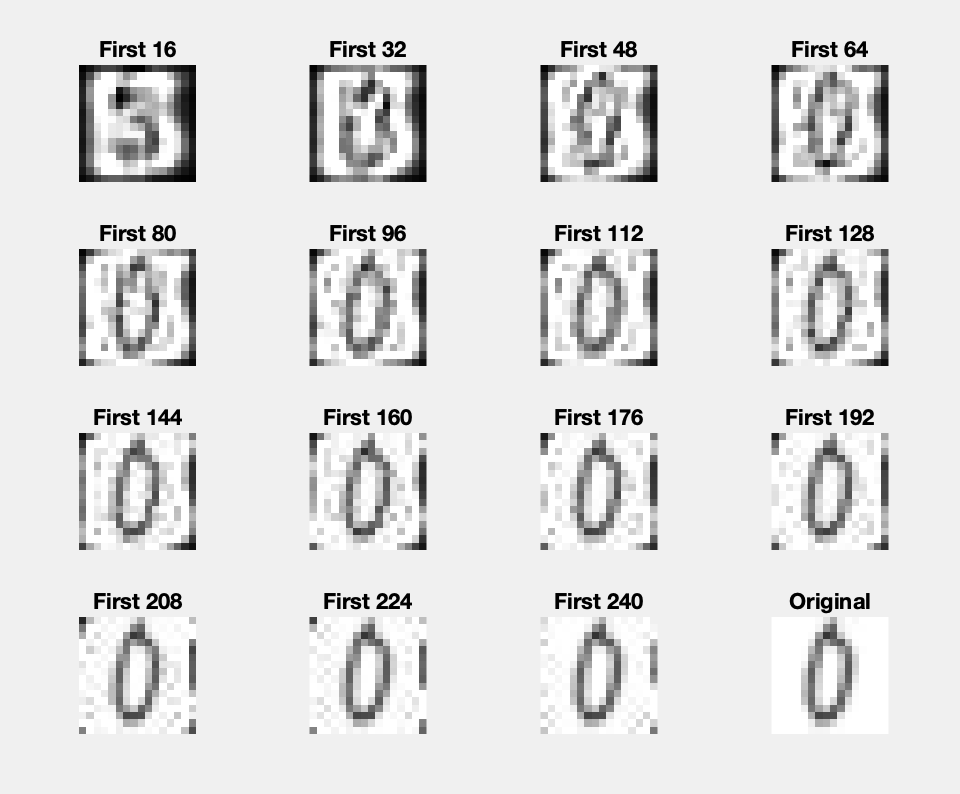
1. By observing these plots, we can observe that the energy distribution of the original data is relatively even without significant energy concentration on any specific dimensions. In the energy distribution plot after KLT, the energy distribution of ST is concentrated in the first few components, which is because ST reflects the overall scatter of the dataset, and the first few components usually contain most of the information. In the energy distribution plot after KLT, the energy distribution of SB is concentrated in the later components, which is because SB reflects the discriminative power between different classes, and the later components usually contain more class-discriminative information.



In the 2-D plot, we can see that there is some overlap between different classes of handwritten digits, they are relatively difficult to distinguish. The separation of the digit 0 is relatively low, as its data points in 2-D space have lots of overlapping. In contrast, the separation of digit 1 and digit 4 are relatively high, as their data points in 2-D space are more scattered and less likely to be confused. Overall, the separability of the handwritten digit dataset in 2-D space is acceptable.



In the 3-D plot, we can see that the separation between different digits is more evident, and the data points of different digits are more scattered. The separation of digit 1 and digit 4 is higher in the 3-D plot, as their data points in 3-D space are more scattered and less likely to be confused. The separation of digit 0 is also improved, as their data points in 3-D space have less overlap. Overall, the separability of the handwritten digit dataset in 3-D space is better, with more evident separation between different digits.



We can see the last plot (original digit) is 0. And easy to observe that the first plot containing 16 eigenvalues seems like 3 instead of 0, therefore, it is not distinguishable when only keeping 16 eigenvalues. However, we can see that the second plot is 0, thus, it is enough to keep 32 eigenvalues for keeping the original digit distinguishable.