BLG 454E Learning From Data (Spring 2018)

Homework 3

1 Question 1

1.a Why to Use Dimension Reduction?

It is useful to use dimension reduction to handle data with large dimensionality since It reduces the size of the data. Moreover, some of the dimensions could be correlated to other dimensions; hence, they are redundant to use.

Another motivation to use dimension reduction is to make data more easy to visualize in small dimensions. It is hard to visualize multidimensional data.

1.b Performance of Dimension Reduction

PCA uses mean and variance to select only dominant features. This is useful to remove noise from data by eliminating redundant or highly correlated features. However, this may not be the case for classification since some datasets cannot be classified using only variance and mean information.

By using this information we can say that performance of a dimension reduction algorithm can be evaluated by considering if the algorithm works well for both linear and non-linear datasets, how it filters noise, efficiency for large datasets, and ease of use in visualization.

1.c Performance of PCA Classification for Given Figure

The performance of PCA in terms of classification is pretty bad for the given figure. Projection shows that data points of different classes are highly mixed up. The accuracy of the classification would be very low for this situation. This is a drawback of PCA which is discussed in the following section.

1.d Drawbacks of PCA

PCA highly depends on data to be linear. The major drawback relies on the linearity assumption of PCA. If the data is not linearly correlated, then the results will not be accurate.

For performance wise, PCA is not efficient for large datasets with high dimensions because of the covariance matrix calculations.

1.e PCA Plot

PCA is implemented in MATLAB language. The dataset contains 64 features and PCA projects data on the plane of 2 features with largest variance. In the Figure 1, only the labels of randomly selected 200 data points are shown. The labels are mostly well grouped as it can be seen from the figure.

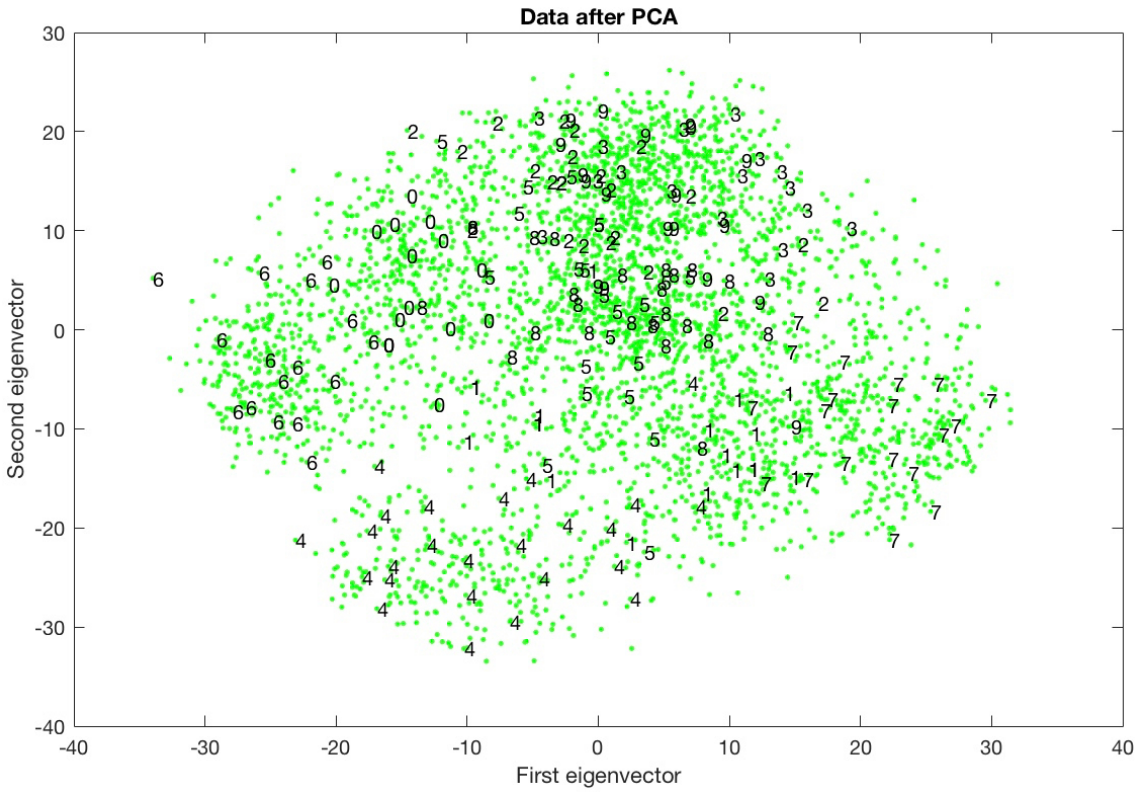

Figure 1: Data plotted in the space of two principal components

2 Question 2

The original image and compressed images are given in the Figure 2. Clustered SVD compresses the image by just using specified number of variances called as rank. In this homework, we are asked to use rank values of 1, 5, 20, 50 for SVD approximation.

It can be seen from the Figure 2 that when such a small number (compared to size of the image) like 50 is used as rank, we obtain a significantly detailed image. Furthermore, increasing the value of rank will increase the quality of the image but even the rank value of 50 can show detailed image.

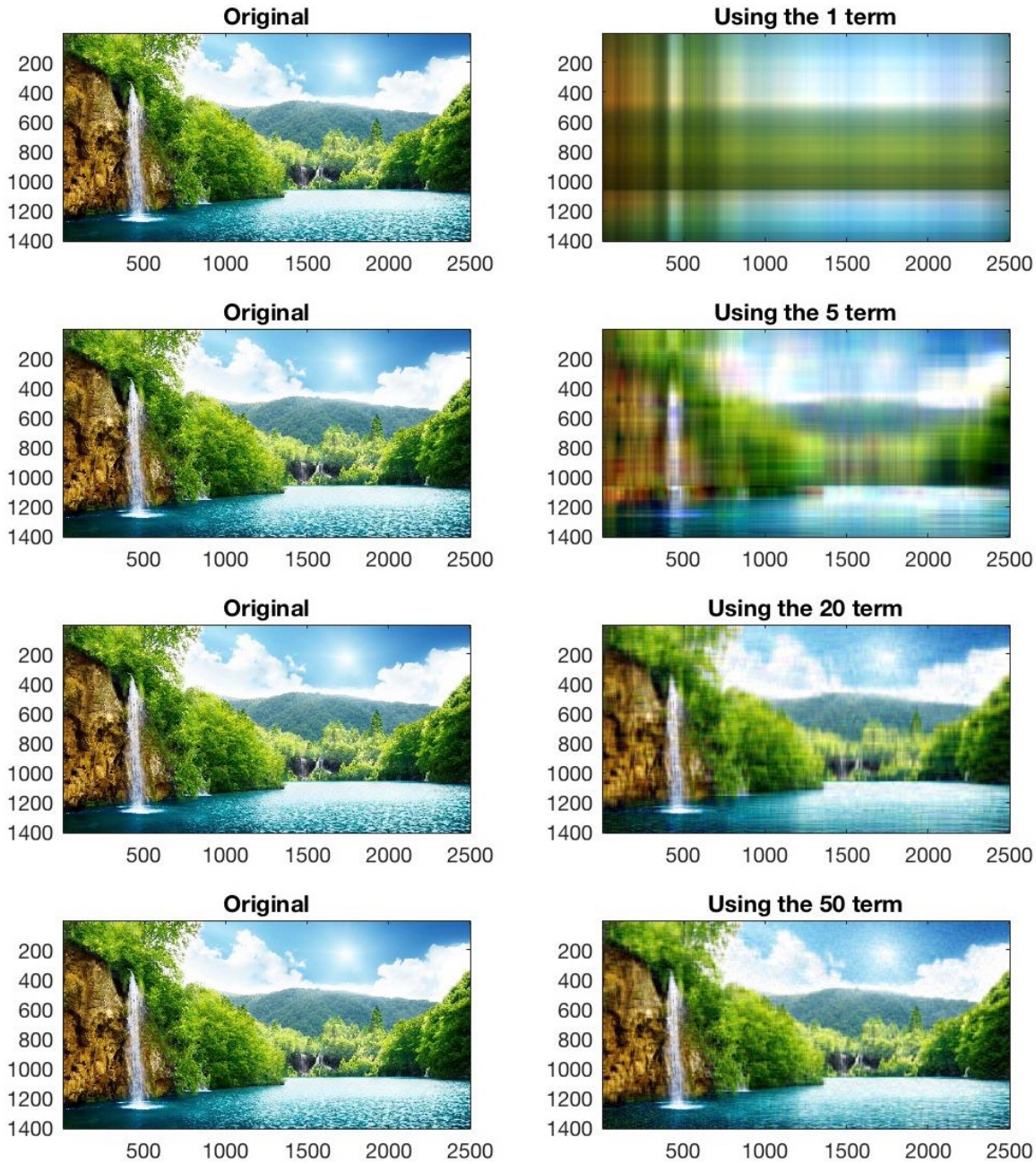

Figure 2: The original image and its the compressed results are displayed