IE-452/IE-552: Algebraic and Geometric Methods in Data Analysis
Course Project-II

Due Date: 10 May 2020

1. **Description**: Download the data file from the course website. The data set contains 5,000 rows to represent the numbers and their labels. For this assignment, Σ and P matrices can be found on the course webpage.

   - Run PCA on the dataset. Plot the eigenvalues in descending order. How many components would you choose by examining this plot only?
   - Examine the first 2 principal components. Create a scatter plot with each of the rows of the dataset projected onto the first two principal components (Your plot must use a different color for each population and include legend). Comment on the facts about the plot and interpret the first two principal components.
   - Examine the third principal component of X. Create a scatter plot with each individual projected onto the subspace spanned by the first and third principal components. What information does the third principal component capture?
   - Display the sample mean for the data set as an image and display the bases which you chose as images. Comment on the covariance matrix.
   - Choose **at least** 10 subspaces with different dimensions and project the data.

2. Assume that d-dimensional zero-mean random vector X with covariance matrix $\Sigma \in \mathbb{S}_+^{d\times d}$,

   - Find the eigengap and maximum eigenvector.
   - Consider $P \in S^{d\times d}$, find the unique maximal eigenvector of the perturbed matrix $\hat{\Sigma} = \Sigma + P$.
   - For different fixed dimensions r=$\{10, 50, 100\}$, find the optimal subspace and the reconstruction error of Σ for each dimension on this r-rank projection.