# Statistical Learning for Engineers (EN.530.641)
# Homework 7

Jin Seob Kim, Ph.D.
Senior Lecturer, ME Dept, LCSR, JHU

Out: 11/04/2022
due: 11/11/2022 by midnight EST



**1** In this problem, you will perform image compression which is basically principal component analysis (PCA). First download a dataset in Scikit-Learn by:
`from sklearn.datasets import fetch_openml`
The data MNIST is about hand-written zip codes. Then load data by:
`X, y = fetch_openml('mnist_784', return_X_y=True)`
Then standardize the data by using `StandardScaler` in `sklearn.preprocessing`.
Now apply PCA. You will have to use `PCA` from `sklearn.decomposition`. Set the number of principal components `n_components` by 0.95, which mean only 95% of the variance will be preserved.

(a) Generate the plot of "cumulative explained variance" vs. dimensions (or number of principal components). To do that, you will have to use `cumsum` from `numpy` with `pca.explained_variance_ratio` as its input.

(b) From the plot, what is the number of principal components (i.e., dimensions) corresponding to preserving 95% cumulative variance?

(c) Select one image randomly from the original dataset, and plot "before" and "after" PCA compression.

## Submission Guideline

- Submit your homework answers in a single pdf format, including plots and so on, to "HW7_analytical" on the gradescope.
- Submit all your python codes in a single .zip file that contains codes for each problem (name them by including the problem number). Name your single zip file submission as "YourName_HW7.zip". For example, "JinSeobKim_HW7.zip" for a single zip file. Submission will be done through "HW7_computational" on the gradescope.

- Just in case you have related separate files, please make sure to include *all the necessary files*. If TAs try to run your function and it does not run, then your submission will have a significant points deduction.
- Make as much comments as possible so that the TAs can easily read your codes.