

PCA and LDA comparison.

The task for this exercise is to compare PCA and LDA in their ability to cluster when projecting very high-dimensional datapoints to 2 or 3 dimensions. In particular, consider the dataset MNIST provided on Virtuale. This dataset contains images of handwritten digits with dimension 28×28 , together with a number from 0 to 9 representing the label. You are asked to:

- Load the dataset in memory and explore its **head** and **shape** to understand how the informations are placed inside of it;
- Split the dataset into the X matrix of dimension $d \times N$, with $d = 784$ being the dimension of each datum, N is the number of datapoints, and $Y \in \mathbb{R}^N$ containing the corresponding labels;
- Choose a number of digits (for example, 0, 6 and 9) and extract from X and Y the sub-dataset containing only the considered digits. Re-call X and Y those datasets, since the originals are not required anymore;
- Set $N_{train} < N$ and randomly sample a training set with N_{train} datapoints from X (and the corresponding Y). Call them X_{train} and Y_{train} . Everything else is the test set. Call it X_{test} and Y_{test} .
- Implement the algorithms computing the PCA and LDA of X_{train} with a fixed k . Visualize the results (for $k = 2$) and the position of the centroid of each cluster;
- For both the algorithms, compute for each cluster the average distance from the centroid. Comment the result; [what should be the centroid? centroid for each cluster or the centroid of all the data](#)
- For both the algorithms, compute for each cluster the average distance from the centroid on the test set. Comment the results;
- Define a classification algorithm in this way: given a new observation x , compute the distance between x and each cluster centroid. Assign x to the class corresponding the the closer centroid. Compute the accuracy of this algorithm on the test set and compute its accuracy for both PCA and LDA;
- Repeat this experiment for different values of k and different digits. What do you observe?

Visualizing dyad.

Consider an image from `skimage.data`. For simplicity, say that $X \in \mathbb{R}^{m \times n}$ is the matrix representing that image. You are asked to visualize the dyad of the SVD Decomposition of X and the result of compressing the image via SVD. In particular:

- Load the image into memory and compute its SVD;
- Visualize some of the dyad $\sigma_i u_i v_i^T$ of this decomposition. What do you notice?
- Plot the singular values of X . Do you note something?

- Visualize the k -rank approximation of X for different values of k . What do you observe?
- Compute and plot the approximation error $\|X - X_k\|$ for increasing values of k , where X_k is the k -rank approximation of X .
- Plot the compression factor $\frac{k}{mn}$ for increasing k ;