# Question 4
# Principal Component Analysis (PCA)

Atishay Jain
Gohil Megh Hiteshkumar

October 15, 2022

## 1 Idea

For given question we can say that,
for a given digit each pixel is a random variable and each instance of those images corresponding to that digit are the values drawn from the random variable in 784 dimensional space.
That is there are total of 784 random variables for a digit.
So,

- to calculate mean($\mu$) for each digit is equivalent to calculating mean for each pixel(mean of each random variable).
- to calculate covariance matrix($C$) means to calculate covariance of two pixels(two RV) and insert it to matrix.
- Since covariance matrix is a $784 \times 784$ real valued symmetric matrix, then according to spectral theorem it has 784 real eigenvalues and 784 real valued eigenvectors.
- to calculate principal mode of variation, first get the max of all eigenvalues then get corresponding eigenvector.

## 2 Implementation

First convert every $28 \times 28$ integer data to floating-point data.
Then reshape each $28 \times 28$ matrix data to a column vector of size 784.

### 2.1 To Get mean Image of every digit

- Create a mean matrix with 784 rows and 10 columns
- each column correspond to a digit (column 1 - digit 0; column 2 - digit - 1)

- we have been given labels of each image in vector for, use it to identify image and add the column vector corresponding to that image in the appropriate column vector of mean image, also we have kept record of total images added corresponding to each mean matrix.

- after all the summation, divide each mean column vector by corresponding count and get mean Image.

- for e.g., suppose $I_1, I_2, ..., I_N$ are the N images corresponding to digit 1, so mean image of digit 1 will be

$$\mu_1 = \frac{\sum_{i=1}^{N} I_i}{N} \tag{1}$$

$\mu_1$ will be a column vector of size 784, to get image we have to reshape it to $28 \times 28$ matrix and display it.

## 2.2 To get covariance matrix for each digit

- First step is to store all images corresponding to a digit seperately inside a matrix of size $784 \times N$(total images corresponding to a digit). Lets call this matrix $D_i$ where i is a digit.

- We can calculate covariance of two vector given their $\mu$ vector by following formula.

$$C = \frac{(X - \mu)(Y - \mu)^T}{N} \tag{2}$$

- By above method we can calculate each entry of covariance matrix of a digit as follows(given below formula calculates entry of covariance matrix for digit 1. Where $X_i$ and $X_j$ are $i^{th}$ and $j^{th}$ column vector of matrix $D_1$)

$$C_n(i,j) = \frac{(X_i - \mu'_{ni})(X_j - \mu'_{nj})^T}{N} \tag{3}$$

where, $X_i$ and $X_j$ are the vectors of size $1 \times N$ and $\mu'_{ni}$ and $\mu'_{nj}$ are also vectors of same size with each entry of these vectors equal to $i^{th}$ and $j^{th}$ value of $\mu_n$ of digit n. We can say that $X_i$ is a RV representing the $(i\%28, i/28)$ pixel.

## 2.3 To get values of $\lambda$(eigenvalue) and eigenvector for calculating principal mode of variation

- Since we have covariance matrix, using eig function in Matlab we can get all eigenvectors and corresponding eigenvalue.

- Now get max of eigenvalues and corresponding eigenvector which will be useful in calculating principal mode of variation.

# 3 Observation

## 3.1 Significant mode of variations

One of the key observation is that the total number of significant modes of variation (i.e., number of "large" eigenvalues) for each digit is very less compared to 784 eigenvalues. Take example of digit 1, the graph of eigenvalue looks like this As we can see here, only 4-6 values are of order $10^5$ and most of other values are around 0. Similar observation can be seen for other eigenvalues graph.
Why this happens,
Since all people write every digit in similar way, then the variation in pixels of these images is small for large no of pixels and it is significant for very less no of pixels. Thus, we get very few significant modes of variation.
Some digits like 4 are written with many different style and hence the number of significant modes of variation for these numbers are quite high than other numbers which are written almost in similar manner by every person, for example 0.

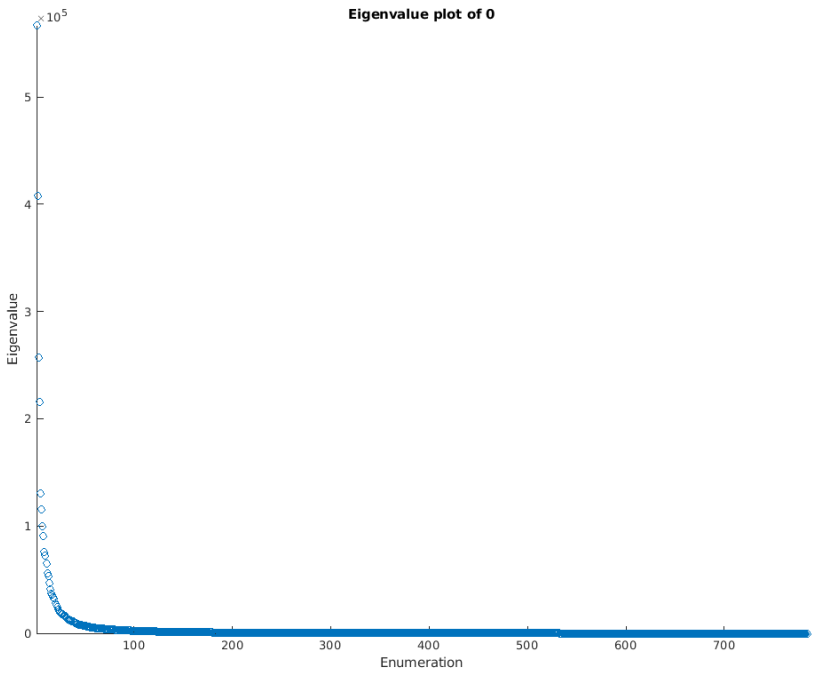


Figure 1: Graph of Eigenvalues of digit 0

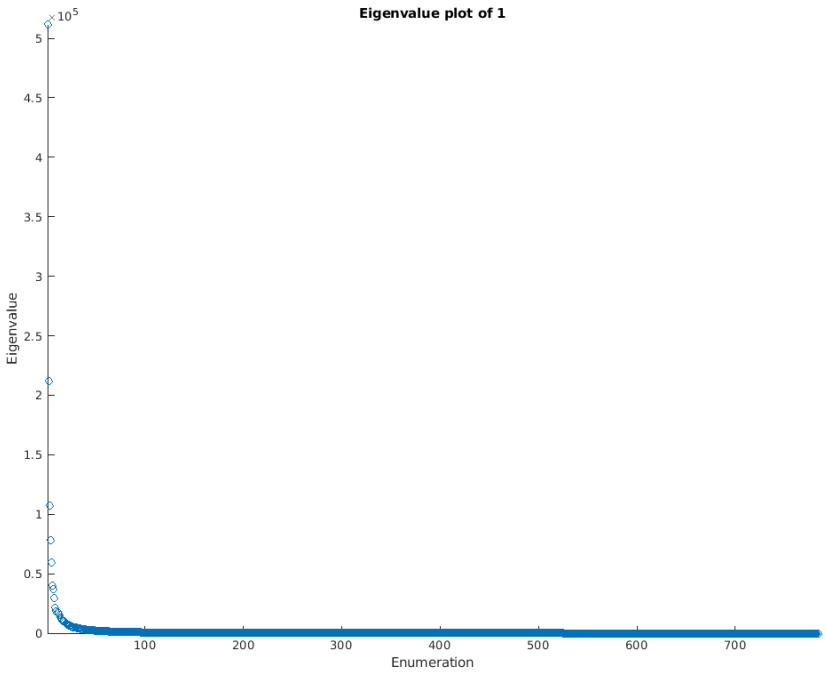


Figure 2: Graph of Eigenvalues of digit 1

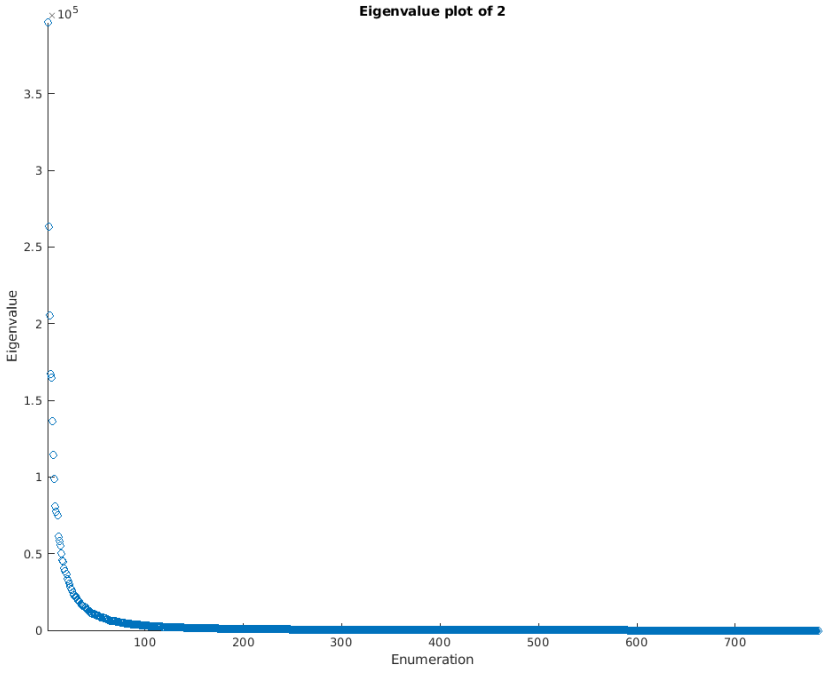


Figure 3: Graph of Eigenvalues of digit 2

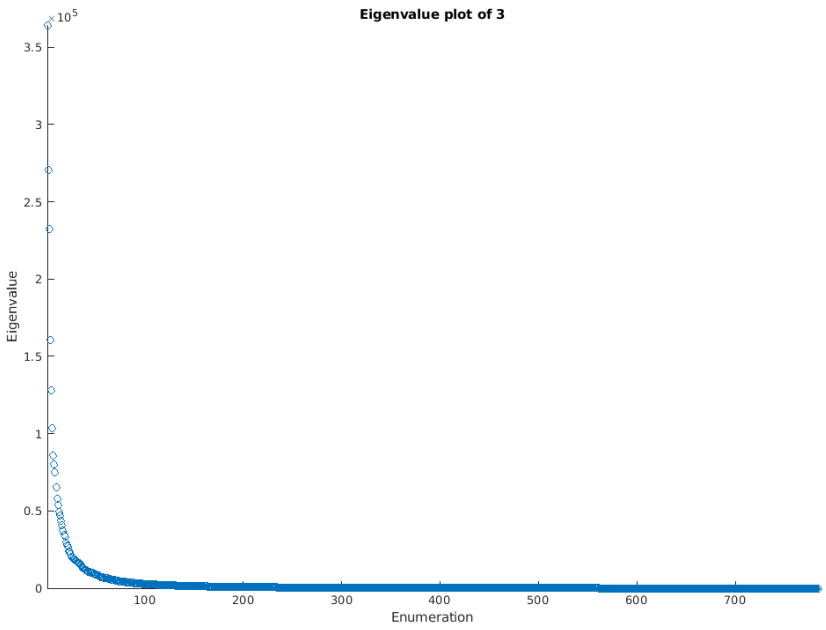


Figure 4: Graph of Eigenvalues of digit 3

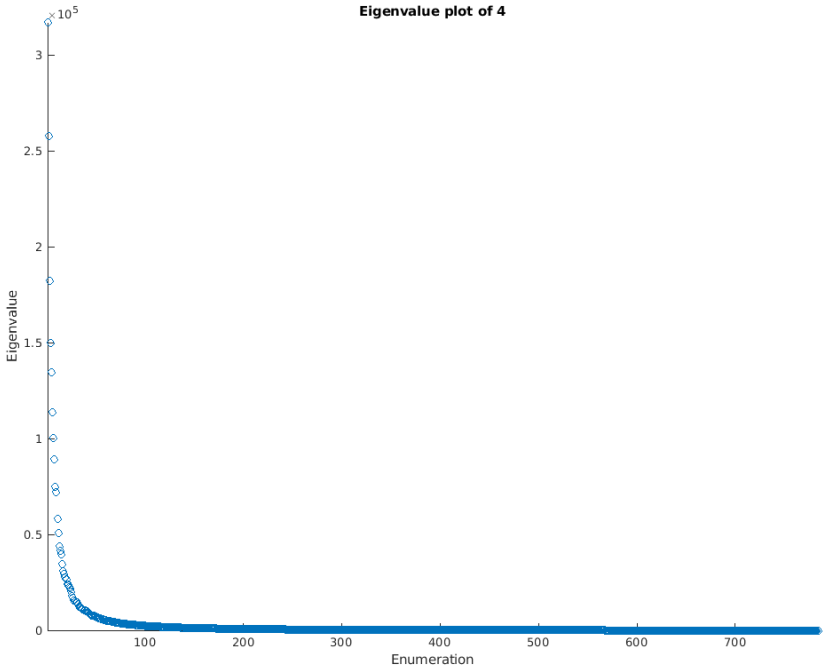


Figure 5: Graph of Eigenvalues of digit 4

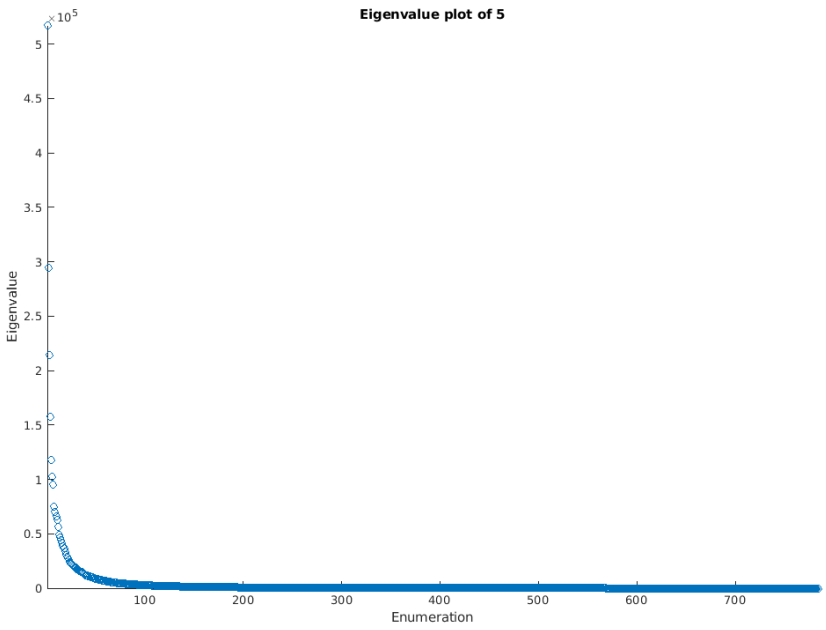


Figure 6: Graph of Eigenvalues of digit 5

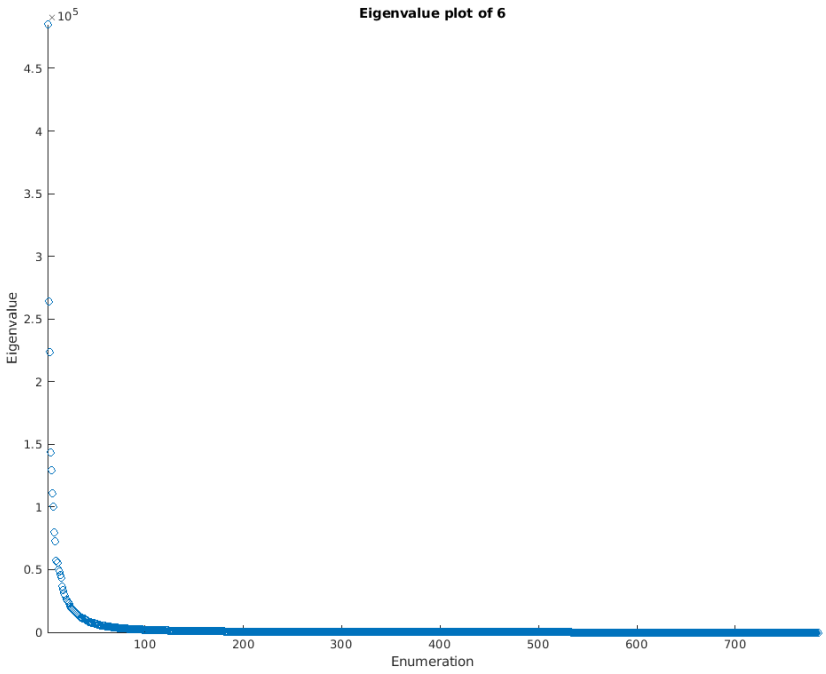


Figure 7: Graph of Eigenvalues of digit 6

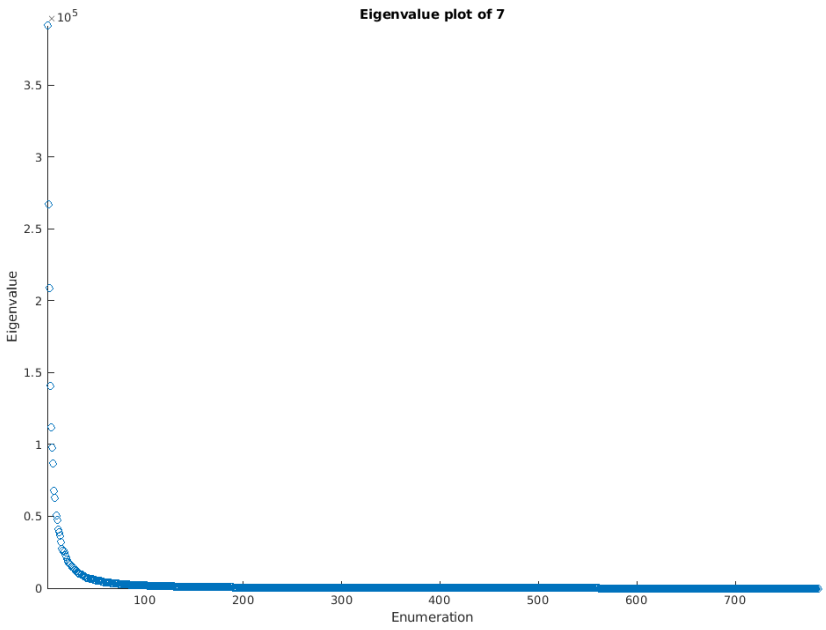


Figure 8: Graph of Eigenvalues of digit 7

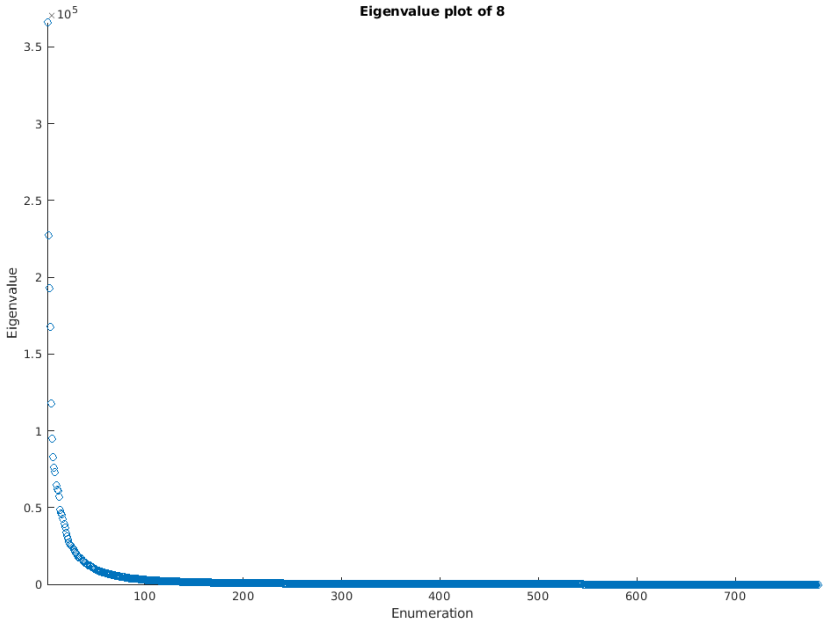


Figure 9: Graph of Eigenvalues of digit 8

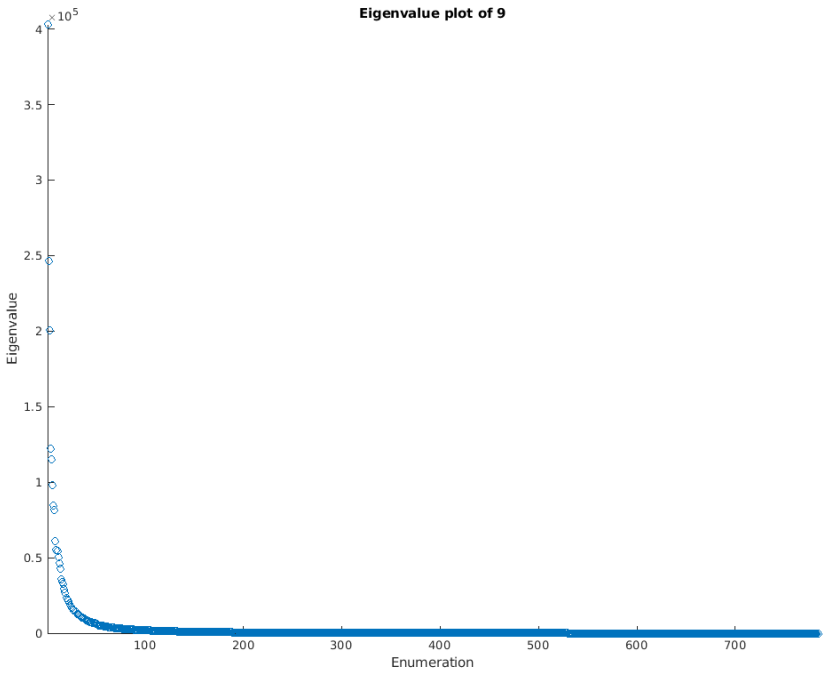


Figure 10: Graph of Eigenvalues of digit 9

## 3.2 Principal mode of variation around mean

As discussed above the pixels with greater variability are less, and to show those variations are depicted by principal mode of variation around mean. So the images generated using these principal mode of variation around mean depicts different styles of writing same digit. Images also show different handwriting style like straight and italic. Lets take example of digit digit 0,
0 can be written in many form, like some people write in circular way and some write it in shape of oval and many more.
So the left image of 0 shows the oval style while right image shows the circular style.
Now talking about italic and straight style, in the pictures we can clearly see one of the left or right image is tilted which shows that people write digits in italic/tilted form. And the other image is almost straight.

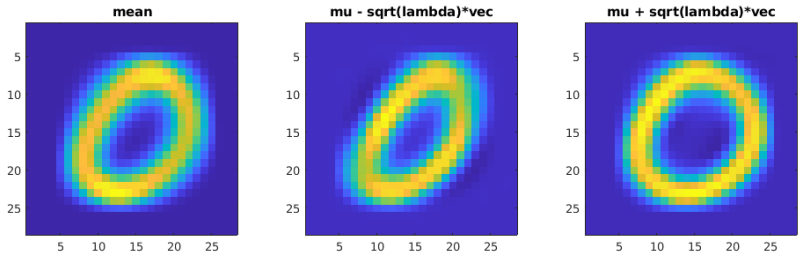


Figure 11: Images for digit 0

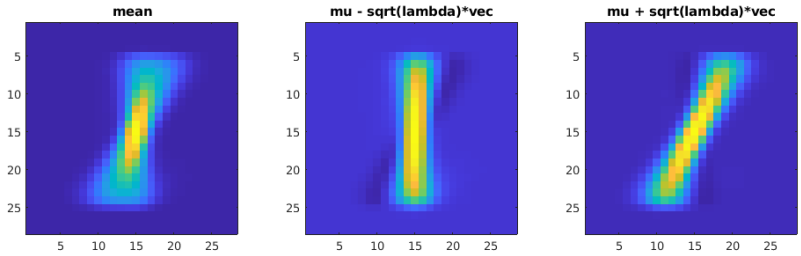


Figure 12: Images for digit 1

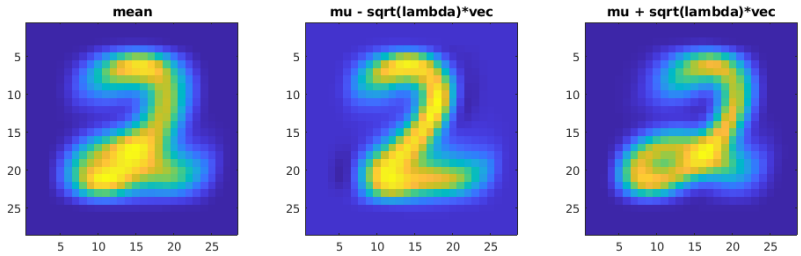


Figure 13: Images for digit 2

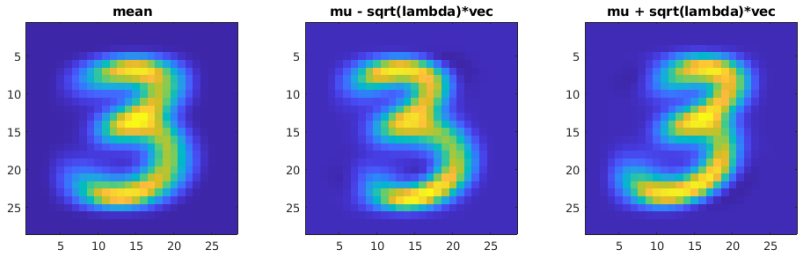


Figure 14: Images for digit 3

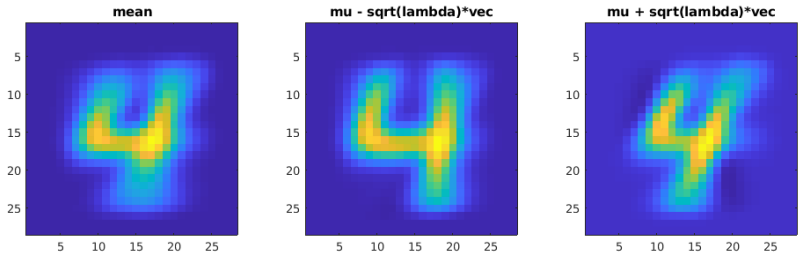


Figure 15: Images for digit 4

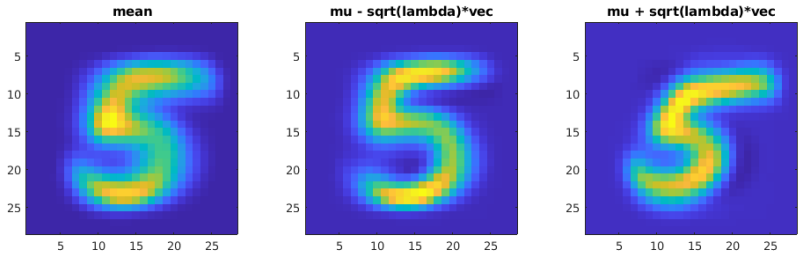


Figure 16: Images for digit 5

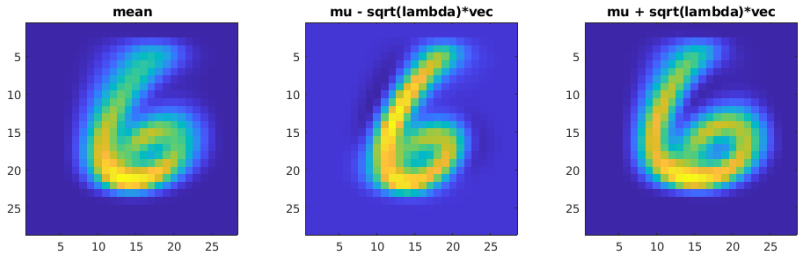


Figure 17: Images for digit 6

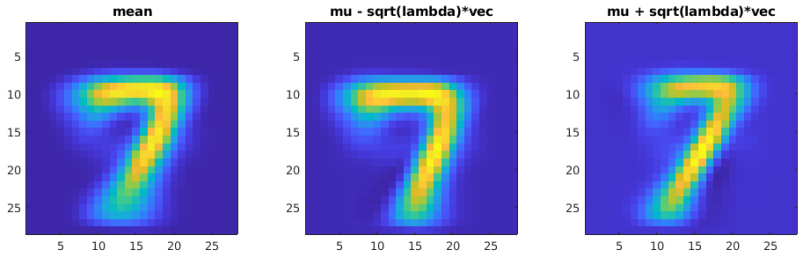


Figure 18: Images for digit 7

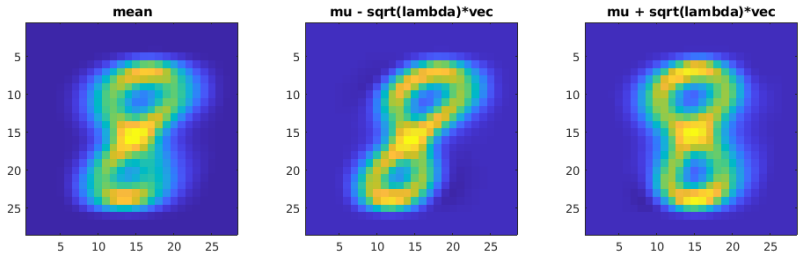


Figure 19: Images for digit 8

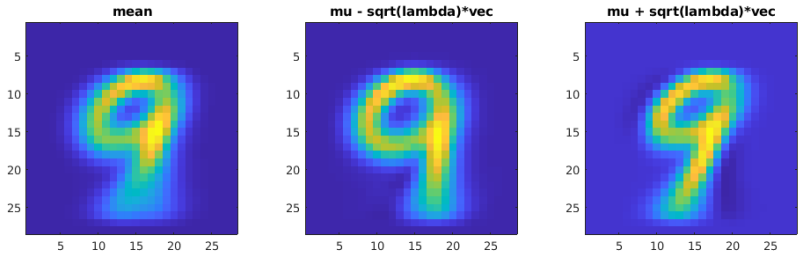


Figure 20: Images for digit 9