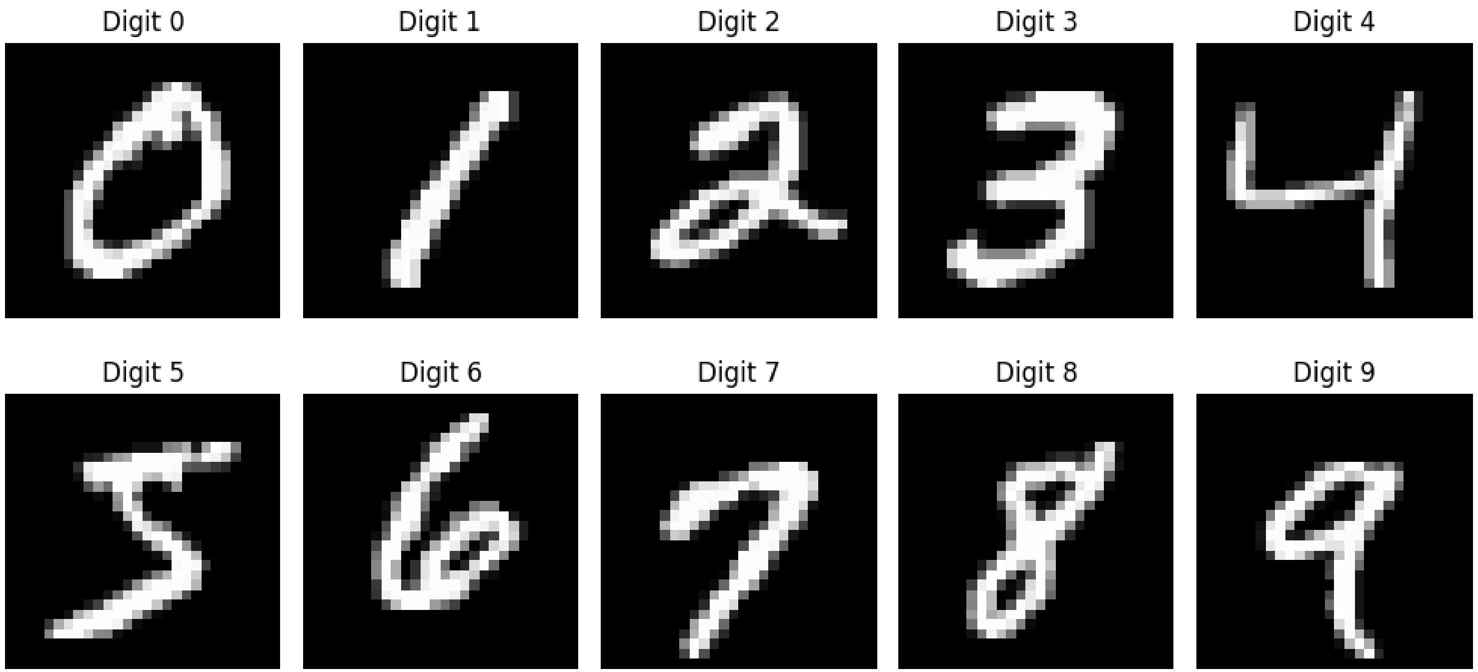
STAT401 – Multivariate Statistical Analysis
2024 Spring

Dimensionality Reduction and Reconstruction of Digit Data

Presenter: Minseo Yoon
(cooki0615@korea.ac.kr)

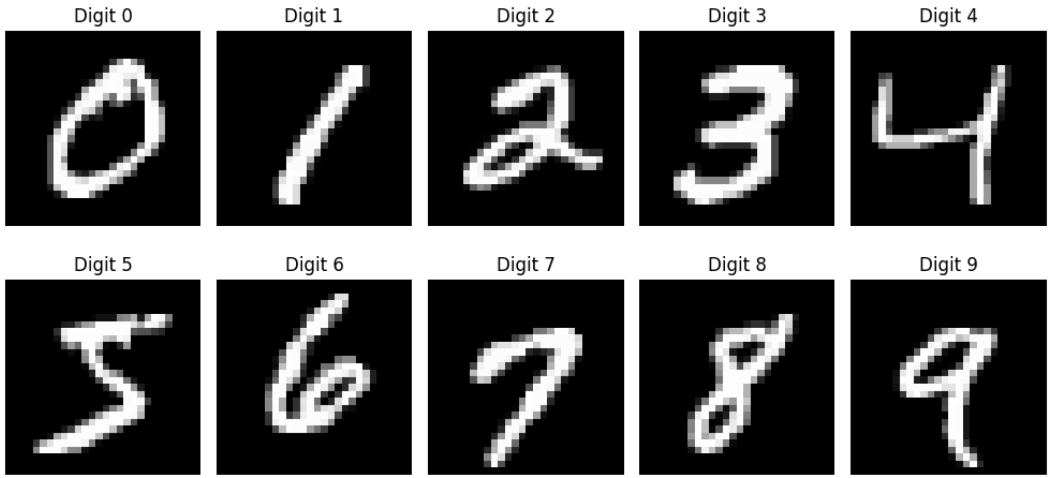
Introduction

- Digit Data (Handwritten)



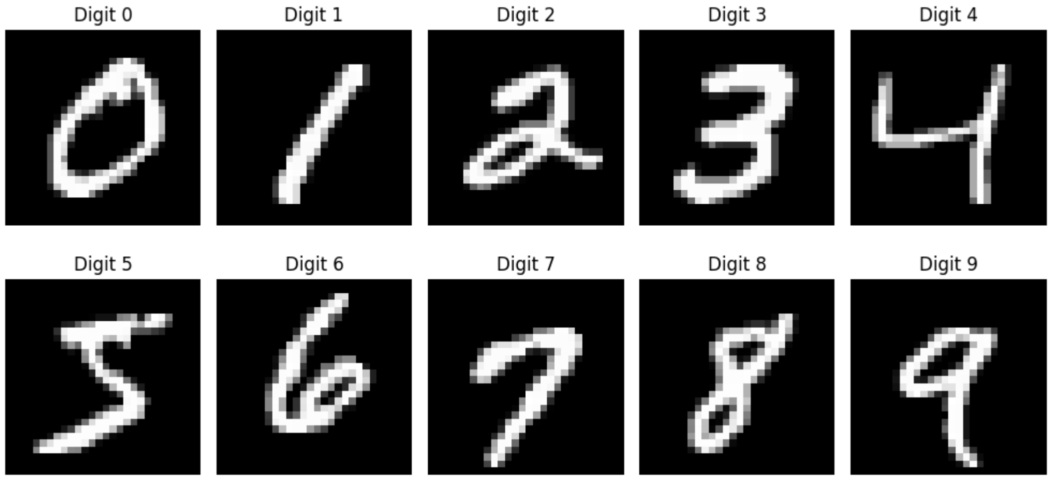
Introduction

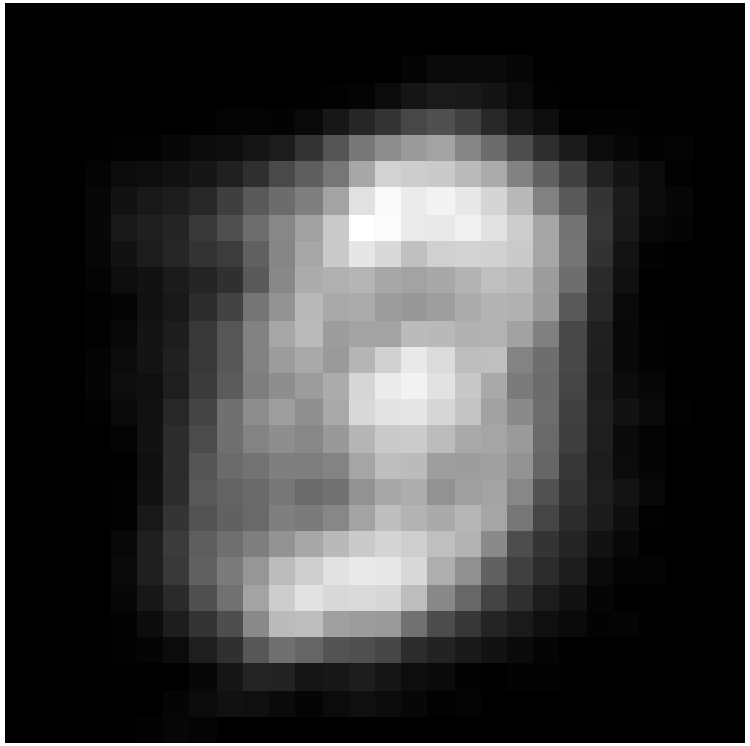
- Digit Data (Handwritten)
  - Name: MNIST
  - 70,000 images
  - 28 x 28 resolution
  - Gray scale (each pixel: 0 ~ 1)



Motivation

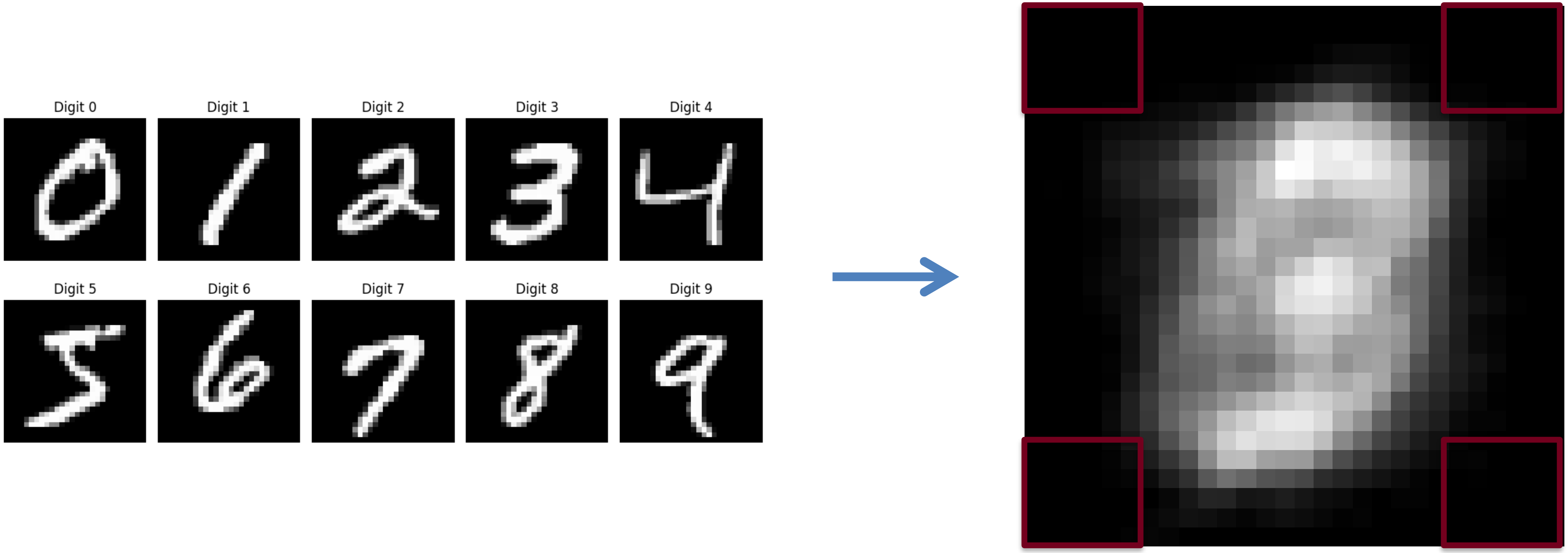
- Why do we need to conduct PCA on digit data?





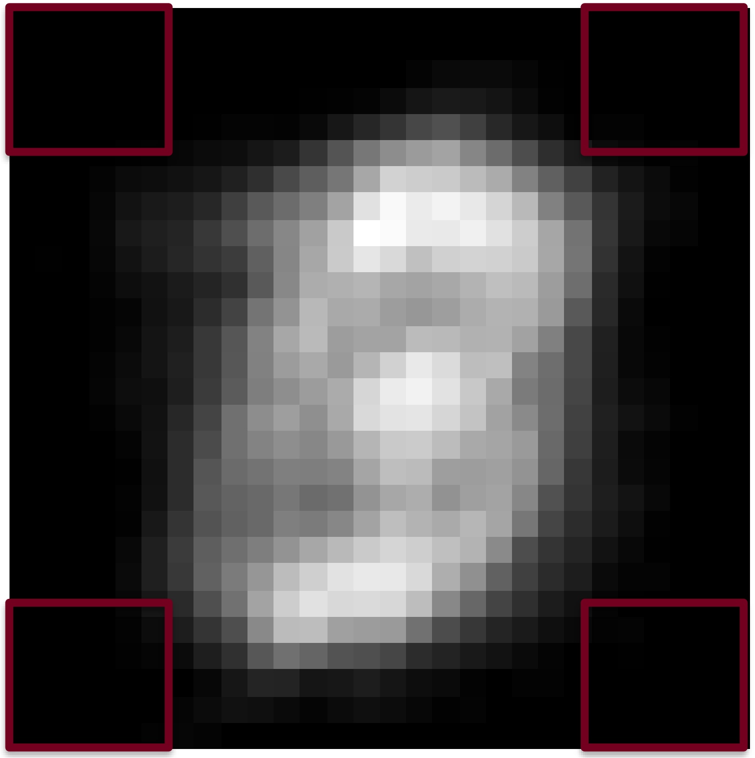
Motivation

- Why do we need to conduct PCA on digit data?



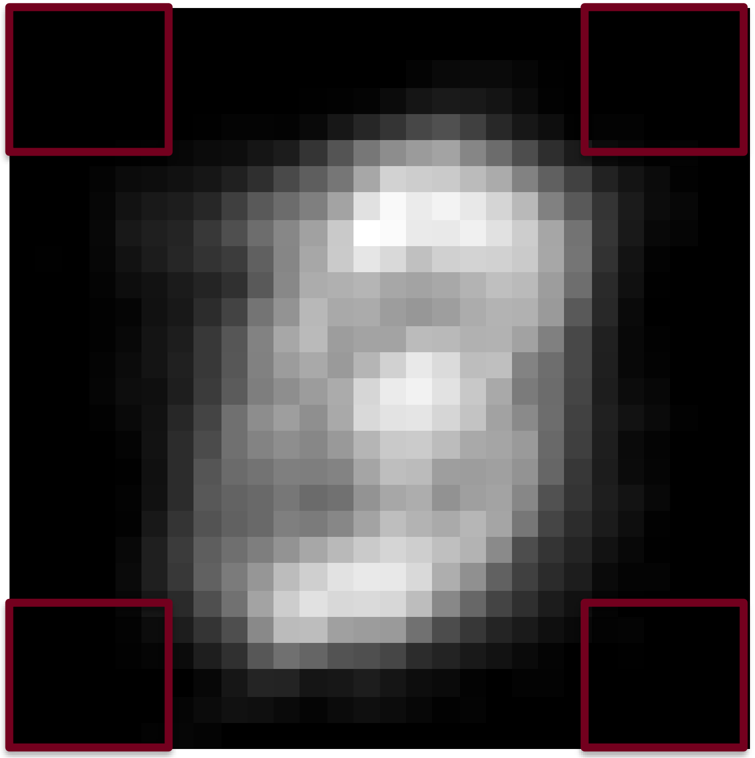
Motivation

- Why do we need to conduct PCA on digit data?
  - There are too many unnecessary pixels!
  - In our words, there are too many redundant features.



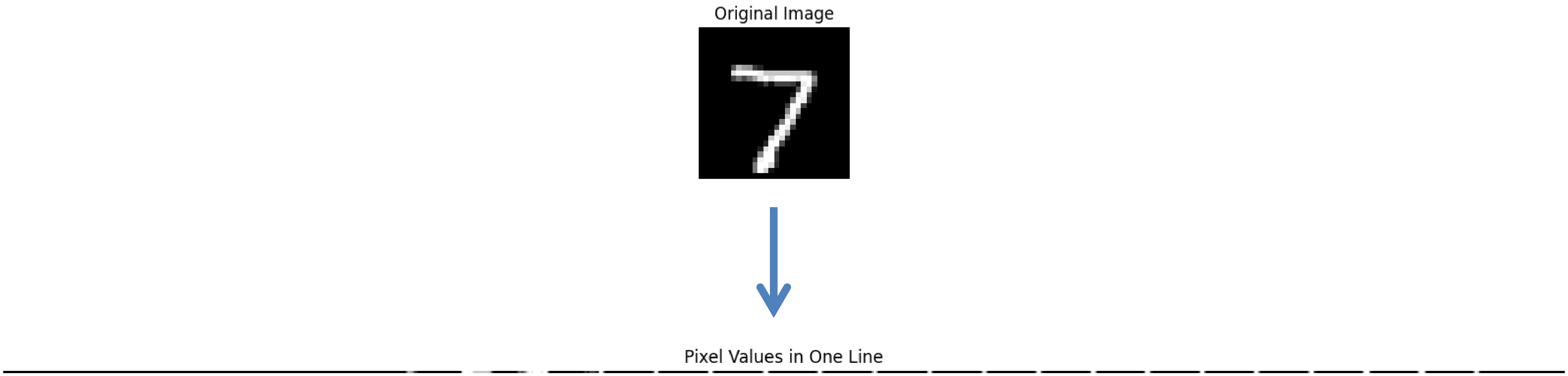
Methodology

- How do we conduct PCA on this data?
  - We have not dealt with image data.
  - This may seem to be more complex than tabular data.
  - But, very simple!



Methodology

- How do we conduct PCA on this data?
  - Just vectorize the matrix-like image data.
  - Then we can simply look at the image as data with 784 variables.



Methodology

- How do we reconstruct the original data using PCA?
  - By Eckart-Young Theorem (Low-rank approximation), we can obtain the best-approximated data only with k principal components.

$$\Sigma = U \Lambda U'$$

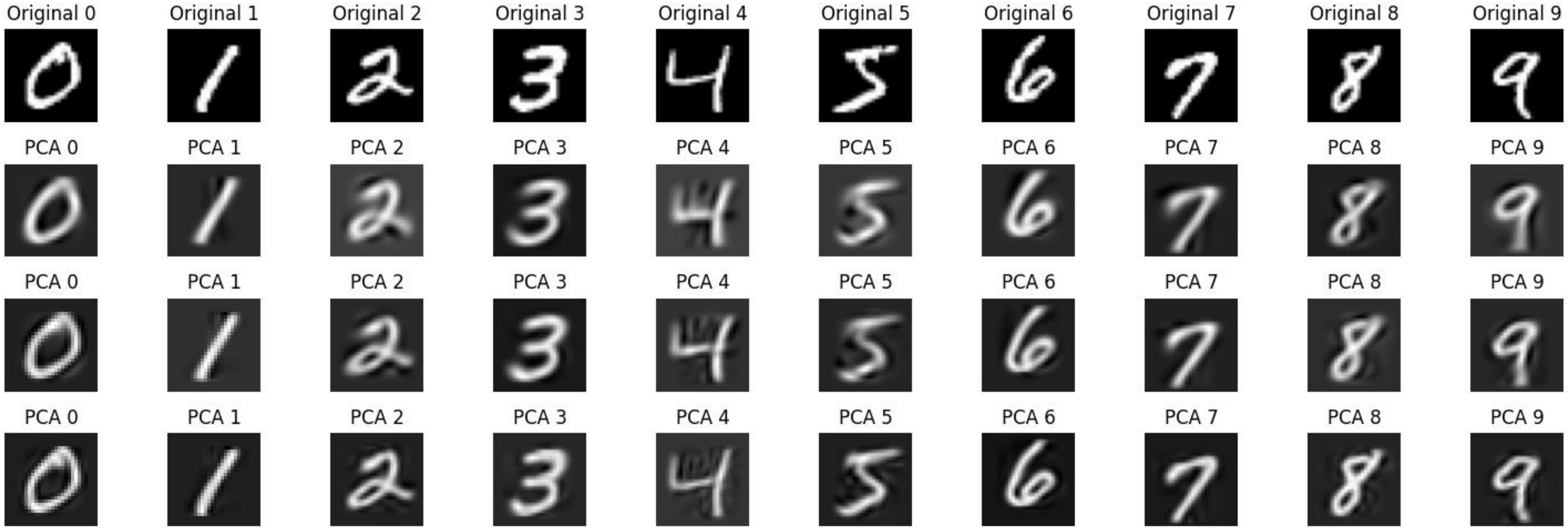
$$X_P = X U_k$$

$$X_R = X_P U_K' = X U_k U_k' \approx X$$

$$\min_{\text{rank}(X_R)=k} ||X - X_R||_F = \sum_{i=k+1}^{28 \times 28} \lambda_i$$

Result and Analysis

- The 2nd, 3rd, and 4th rows represent the results obtained by retaining 70%, 80%, and 90% of the total variations, respectively.

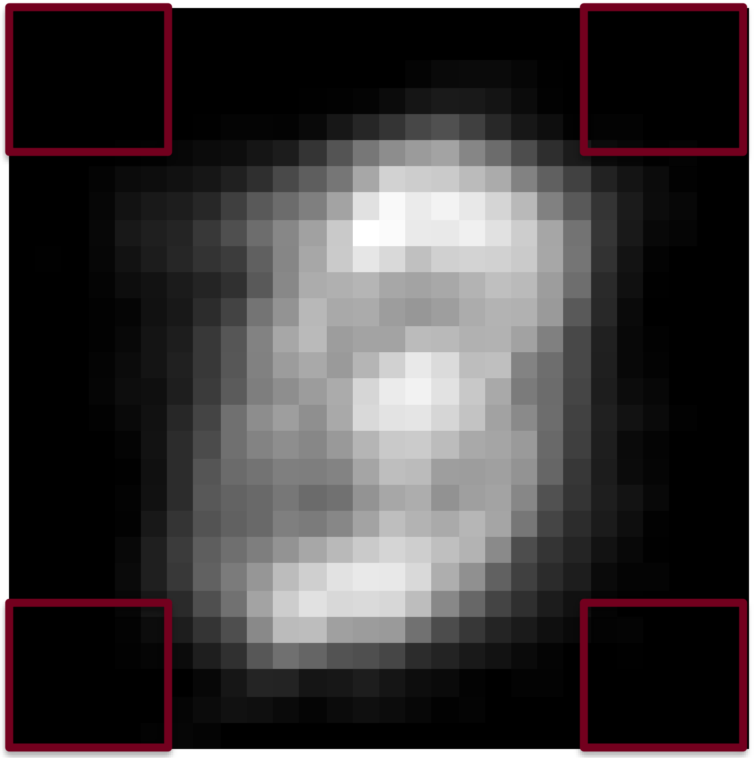


Result and Analysis

- Analysis of the first components

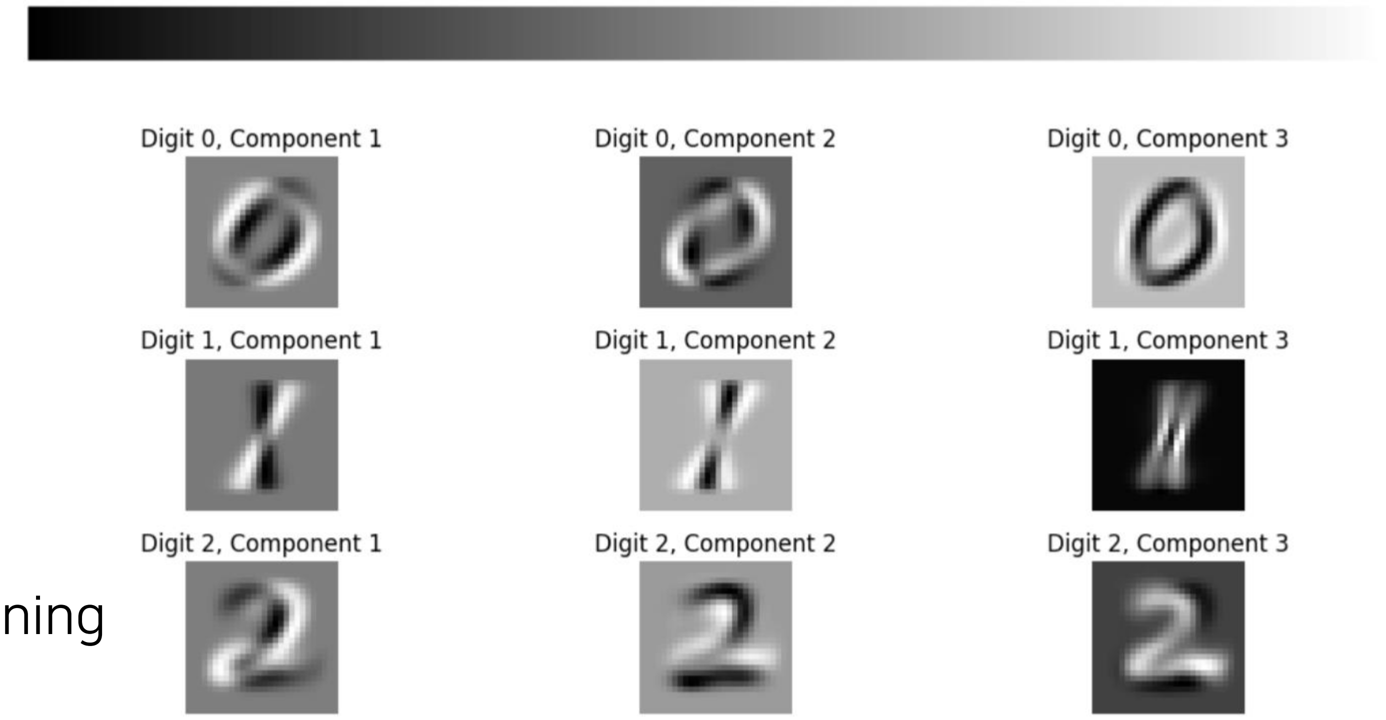
```
Top-left 4x4 component:
[[ 5.26572512e-20 -5.55111512e-17 -5.55111512e-17  0.00000000e+00]
 [ 0.00000000e+00  0.00000000e+00  0.00000000e+00  0.00000000e+00]
 [ 0.00000000e+00  0.00000000e+00  1.96615753e-06  1.09264414e-06]
 [ 0.00000000e+00  0.00000000e+00  2.56757368e-06  7.83497352e-06]]

Top-right 4x4 component:
[[ 0.00000000e+00  0.00000000e+00  0.00000000e+00  0.00000000e+00]
 [ 0.00000000e+00  0.00000000e+00  0.00000000e+00  0.00000000e+00]
 [ 8.82496594e-06  3.65876221e-06  0.00000000e+00  0.00000000e+00]
 [ 1.35708909e-05  1.38381967e-06 -3.22083306e-06  0.00000000e+00]]

Bottom-left 4x4 component:
[[ 0.00000000e+00  0.00000000e+00  4.14579037e-06  1.60971199e-05]
 [ 0.00000000e+00  0.00000000e+00  5.25692968e-07 -7.54683342e-06]
 [ 0.00000000e+00  0.00000000e+00  0.00000000e+00 -9.89068983e-07]
 [ 0.00000000e+00  0.00000000e+00  0.00000000e+00  0.00000000e+00]]

Bottom-right 4x4 component:
[[ 2.75321633e-05 -3.81759275e-06 -2.19428242e-06  0.00000000e+00]
 [ 1.71774789e-05 -7.70769860e-07 -1.59475012e-06  0.00000000e+00]
 [ 6.46282701e-07 -7.60609540e-07  0.00000000e+00  0.00000000e+00]
 [ 0.00000000e+00  0.00000000e+00  0.00000000e+00  0.00000000e+00]]
```

$\approx$



Result and Analysis

- Interpretation of the first three components
  - 1st PC
    - Contrasts right diagonal shape with left diagonal shape
  - 2nd PC
    - Contrasts vertical shape with horizontal shape (maybe)
  - 3rd PC
    - There seems to be hardly any meaning



Future work

- Cluster Analysis
- Discriminant Analysis

❖ The principal components are often used as input for another analysis such as multiple regression and cluster analysis.