Principal Component Analysis (PCA) for Image Reconstruction using MNIST DataSet

Divya Jyoti

(Dated: March 16, 2025)

Abstract

This study explores the application of Principal Component Analysis (PCA) for image compression and reconstruction using the MNIST dataset. PCA is employed to reduce the dimensionality of images while retaining key structural information. The reconstructed images are evaluated using Peak Signal-to-Noise Ratio (PSNR) to measure the fidelity of reconstruction. Results indicate that increasing the number of principal components improves reconstruction quality, highlighting the trade-off between dimensionality reduction and image clarity. This report provides insights into optimizing PCA for image processing tasks.

1. Introduction

This report presents the results of applying Principal Component Analysis (PCA) to reduce the dimensionality of images from the MNIST dataset and reconstruct them using varying numbers of principal components (p-values). The quality of the reconstructed images is assessed using the Peak Signal-to-Noise Ratio (PSNR).

2. Methodology

2.1 PCA Transformation

The dataset consists of handwritten digit images from MNIST, which are transformed using PCA by selecting the top `p` eigenvectors corresponding to the highest eigenvalues. The transformation steps include:

- Centering the images by subtracting the mean vector.
- Projecting them onto the selected eigenvectors.
- Reconstructing the images using the inverse transformation.

2.2 Covariance Matrix Computation

The covariance matrix captures the relationships between different pixel intensities across the dataset. We are using numpy.cov() function from numpy library to compute covariance matrix.

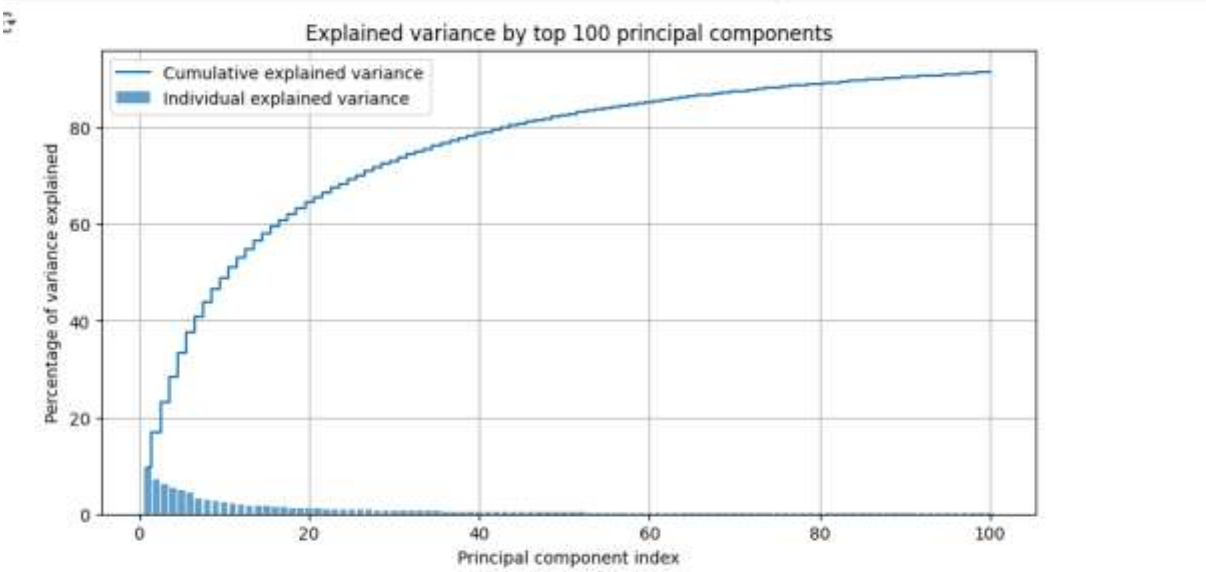
2.2 Explained Variance and Cumulative Variance

PCA finds a new set of orthogonal basis vectors (principal components) that maximize variance in the data. The amount of variance explained by each principal component is given by:

$$\text{Explained Variance} = \frac{\lambda_i}{\sum \lambda}$$

where λ_i is the eigenvalue of the i-th principal component. The cumulative explained variance is calculated as:

$$\text{Cumulative Explained Variance} = \sum_{i=1}^{p} \text{Explained Variance}_i$$



This metric helps determine how many principal components are needed to retain a significant portion of the original data's variance.

2.3 PSNR Calculation

PSNR is calculated to quantify the reconstruction quality:

PSNR= 20 * np.log10(max_pixel / np.sqrt(mse))

where Mean Squared Error (MSE) is the average squared difference between the original and reconstructed images.

3. Results

3.1 PSNR Values for Different p-values

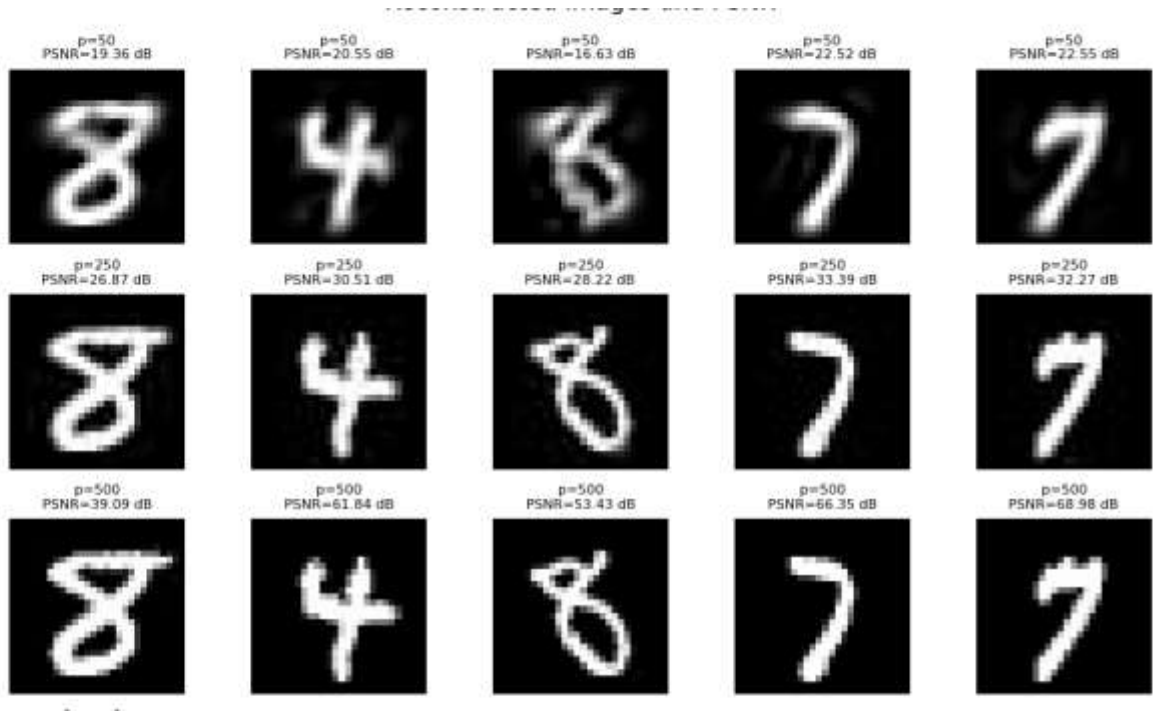
The PSNR values for different values of `p` (number of principal components) are summarized below:

p-value	Image 1	Image 2	Image 3	Image 4	Image 5
50	19.36 dB	20.55 dB	16.63 dB	22.52 dB	22.55 dB
250	26.87 dB	30.51 dB	28.22 dB	33.39 dB	32.27 dB
500	39.09 dB	61.84 dB	53.43 dB	66.35 dB	68.98 dB

3.2 Visual Representation

Below is a set of reconstructed images for different values of `p`:

Reconstructed Images and PSNR



4. Observations and Insights

- The PSNR values improve as `p` increases, which is expected as more principal components capture more image details.
- The lower PSNR values for smaller `p` indicate higher reconstruction error due to loss of important information.
- The significant PSNR jump at `p = 500` suggests that a large portion of the image variance is captured by higher principal components.

5. Reference:

https://builtin.com/data-science/step-step-explanation-principal-component-analysis