Quantitative Foundations Project Report 4

Rinaldo Iorizzo, Gursimran Singh

December 2021

1 Part 1 - PCA

1.1 Introduction

Principle component analysis (PCA) is a statistical procedure that uses orthogonal transformation to convert a set of observations of possible correlated variables into a set of linearly uncorrelated variables called principle components. In this project, we have applied PCA in order to reduce the dimension of 400 test images taken from 40 subjects of a well-known face database. Then, taking images from different subjects, we reconstructed the given images using a different number of principle components (so-called eigenfaces). By using a different number of principal components, it can be shown that effective dimension reduction of the input can be achieved by observing images from the reduced space and their apparent similarity to the original image. Further analysis of the eigenfaces themselves can be insightful, as interpretation of these images may inform us of the importance of various qualities in our data that would have been hard to quantify previously.

1.2 Dimension Reduction

Dimension reduction can be illustrated by the visualization of eigenfaces (figure 1 produced through PCA. In viewing the eigenfaces, we take note of the apparent structure in each image. As the eigenfaces are ordered according to the variance they account for, the leading eigenfaces represent the portion of each image that strongly indicates the structure of a human face. The first seems to indicate the outline of a human face, while the second highlights the hair and hairline. The third seems to show the sides of the head and neck. Further eigenfaces are increasingly hard to interpret as obviously meaningful.

1.3 Image Reconstruction

Image reconstruction was performed with a variety of principal components and demonstrated with a variety of images from various subjects. We note that at 100 components (corresponding to roughly 90% of the variance as shown in figure 2) produces an image that resembles a face with noticeable noise. At around 150 components, the image produced is clearer and reconstructs the original with some error. Further increases in the number of components lessens overall noise, but appears to have diminishing returns. We say that between 100 and 150 eigenfaces are required to reconstruct a face with reasonable error.

2 Part 2 - Classification

2.1 Introduction

Classification is a process of categorizing a given set of data into classes, and it can be performed on both structured or unstructured data. The goal of classification is to predict the class of given data points. The classes are often referred to as targets, labels, or categories. Binary classification is a special case when there are only two categories data can fall into, and may be referred to as recognition. In this project, we applied the Multiple Linear Classification model to perform face recognition and face identification on the given image dataset,i.e., binary and multi-class classification on the given facial image dataset.

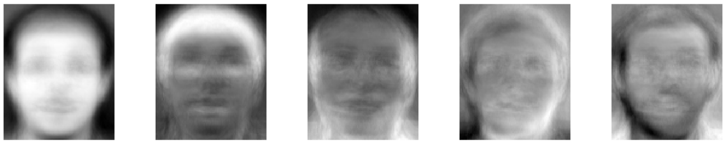

Figure 1: The first five leading eigenfaces, from left to right.

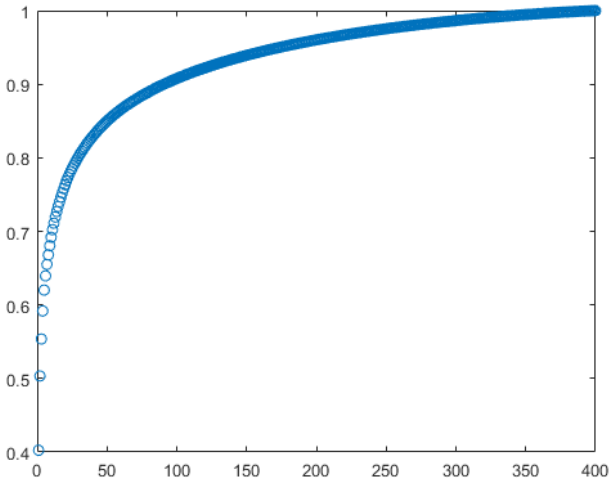

Figure 2: Features contributing to the total variance(most to least)

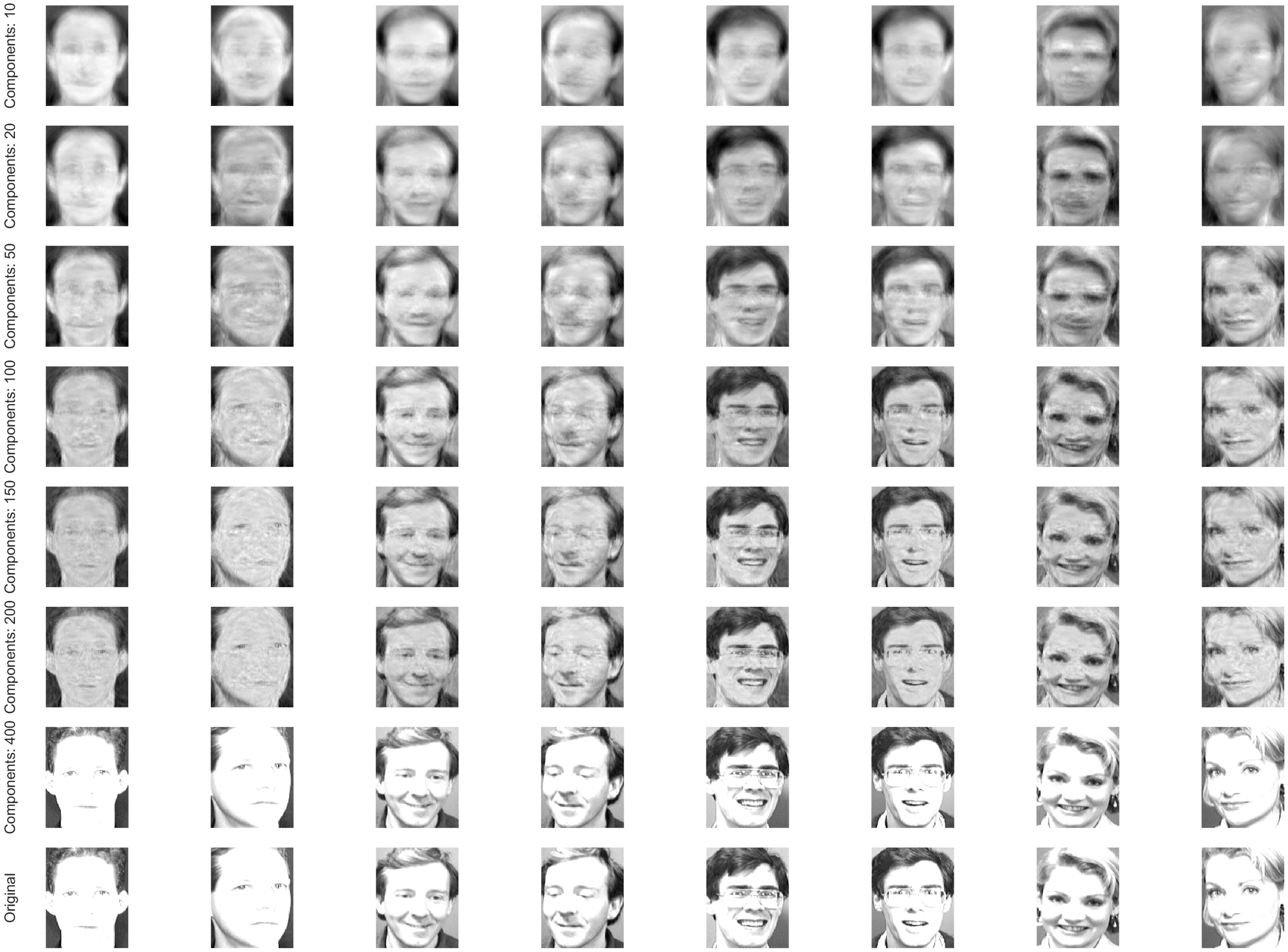

Figure 3: Reconstructed images from the dataset with a variety principal components. Images are from subjects 1 3 20 and 35, images 1 and 5.

2.2 Non-Face Images

Non face images were sourced from a high resolution image that does not contain any human faces. Selections of the appropriate resolution were cropped and saved to produce a set of non-face images.

2.3 Observations and Results

We first perform PCA on our training data and reduce the dimension of our training and test set before classification. For Face Recognition, we achieved an accuracy of 96% with a polynomial basis expansion of degree 1, whereas, for Face Identification, we got achieved an accuracy of approximately 55% with a polynomial basis expansion of degree 2. In both cases, we took the top 100 leading features via PCA for classification. Increases in polynomial rank did not lead to increased accuracy in the test set, for both binary and multi-class classification. An increase in principal components past 150 lead to steep decline in test set accuracy in the multi-class case. We believe our results may be improved by providing a larger set of data samples from all classes are in similar ratios. Additionally, some more advanced machine learning models may improve performance like Logistic Regression, LDA, SVM, etc.