Yale Faces and Music Genre Identification

Peter Schultz - March 6 2019

I. Introduction and Overview

Two applications of the SVD are discussed herein. The first is using the SVD to perform an analysis of images of faces and their major components, and how the components are more important for different faces. The second is using the SVD to deconstruct some different genres of music into their major components, and attempting to use those components to attempt to classify new samples of music. Part 2 uses three genres, Rap, Piano, and R&B.

II. Theoretical Background

As discussed in the previous assignment, the SVD is an extremely powerful tool. Basically it has the ability to break down a matrix X into three parts. This includes **U**, **Σ**, and **V**, following this formula:

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^* .$$

The two components, **U** and **V*** are responsible for the rotation of said vector or matrix, and **Σ** is responsible for its stretching. **Σ** contains the singular values of the matrix, which are related to its the eigenvalues of the matrix A*A or AA* by a square root.

Part 1

Using a dataset of cropped images of assorted peoples' faces provided by Yale, amounting to 2496 images from 39 different people. Across all individuals with their pictures taken, the images have 64 different lighting conditions. This allows us to generate a set of common and important correlated structures that exist within all faces. The second dataset used in this part includes pictures of individuals faces, but these photos are not cropped the same way. Some pictures have the individual's face on one of the sides, or with drastically different lighting conditions. This will cause some problems when attempting to identify correlated principal modes.

Part 2

Using a dataset consisting of various 5 second samples from different genres and artists, we can generate a set of common and important correlated structures, much like what is done with the faces in part 1. By analyzing how each genre projects onto these principal modes, we have the ability to classify new samples based on how they project onto the principal modes generated from a training set. This, however, relies on sufficient separation between the projection profile of the genres in question.

III. Algorithm Implementation and Development

Part 1

To start the analysis of the face dataset, a matrix containing each image is generated, where the columns of the matrix correspond to each of the rows of an image in sequence, and each new column represents a new image. This will serve as the X matrix discussed in Section 1 for this analysis. In this application, we can perform the SVD on the X matrix to obtain its three components. Here, the **U** matrix represents the principal correlated structures (or dominant principal modes) that exist within all of the faces. The **Σ** matrix contains values on the diagonal that represent the relative importance of each principal mode. Lastly, the **V** matrix represents how each face relates to each of the modes.

Part 2

The algorithm section for part 2 is similar to that of part 1. However, the X matrix contains the FFTs of 366 different of 5 second samples of three different genres as the columns, each of the genres with around 100 samples. The 5 second, stereo clips are combined and downsampled by half to save computation time Here, the **U** matrix represents the principal correlated structures (or dominant principal modes) that exist within all of the music samples. The **Σ** matrix contains values on the diagonal that represent the relative importance of each principal mode. Lastly, the **V** matrix represents how each sample relates to each of the modes.

IV. Computational Results

Part 1

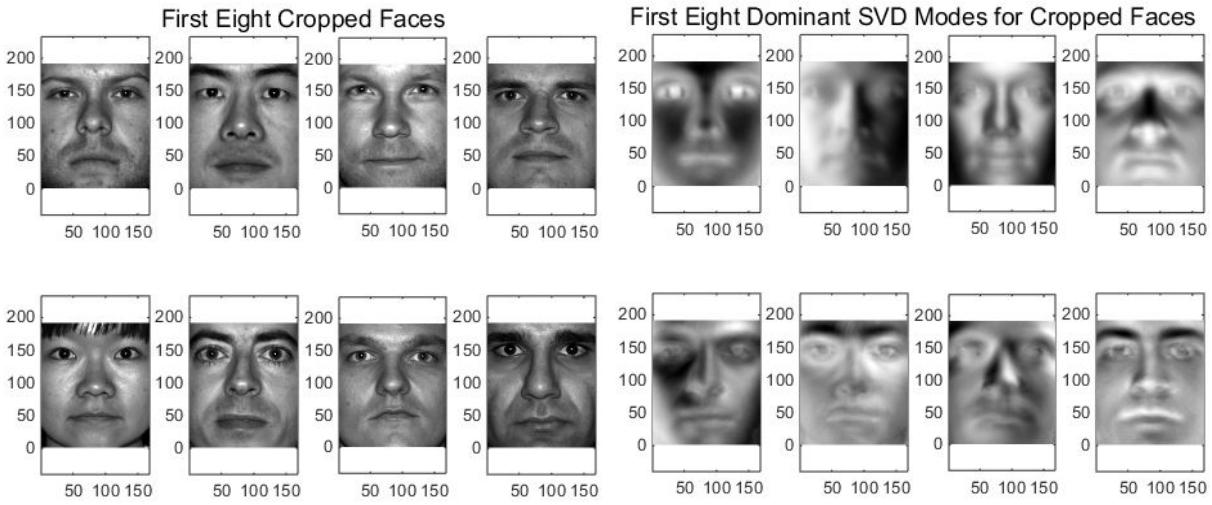

Figure 1: (Left) First eight different faces in Yale dataset. (Right) Eight most important principal modes present in entire Yale dataset.

Figure 1 shows the first eight faces in the Yale dataset and the dominant structures that exist in the dataset. These principal modes are interesting because they highlight the

features of faces that we use to differentiate between people. For example, the top left mode highlights the iris and pupil, while the third mode highlights the bridge of the nose, and general shape of the face.

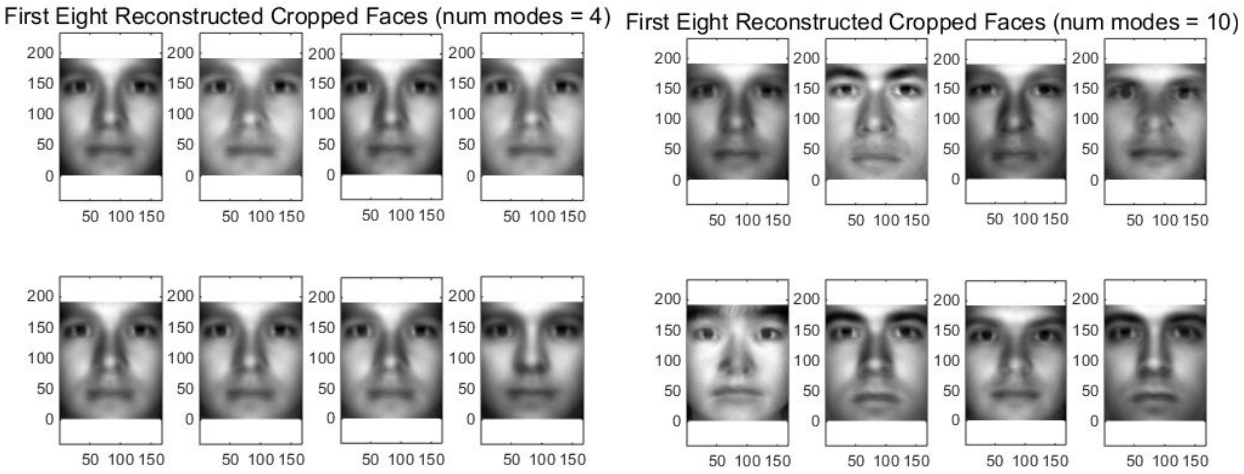

Figure 2: (Left) First eight different faces in Yale dataset reconstructed using four principal modes. (Right) The same with 10 modes.

Figure 2 demonstrates two different reconstructions of the first eight faces in the Yale dataset. The left part of the figure uses only the 4 most important principal modes to reconstruct the faces. It is clear that they are faces, however they appear to be merely the average face in the set. They contain all the normal features such as brows, noses, eyes, lips, but the reconstructions are lacking any specificity to differentiate their identities. The right side of the figure adds in an additional 6 modes up to a total of 10. At this point, the faces begin to be discernible. With this amount of information a person would likely be able to identify the person, but not with perfect accuracy.

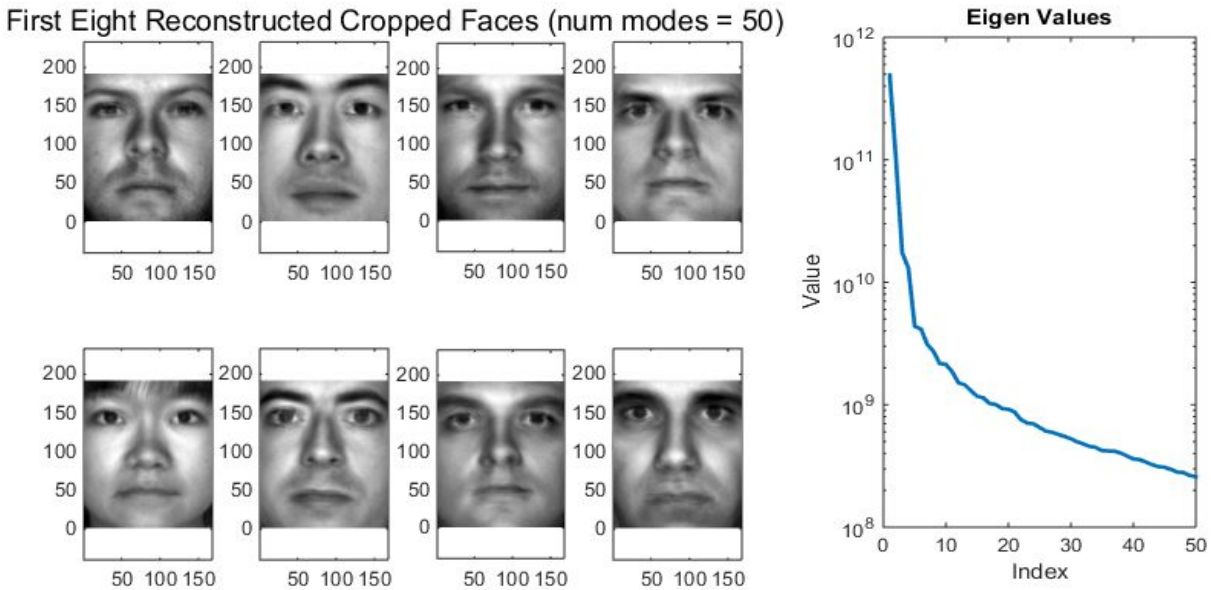

Figure 3: (Left) First eight different faces in Yale dataset reconstructed using fifty principal modes. (Right) Eigenvalues associated with principal modes.

Figure 3 demonstrates similar ideas to figure 2, but with 50 principal modes, and now the faces are clearly recognizable as the first 8 in the dataset. The right portion of the figure shows the steep decline in importance of the principal modes.

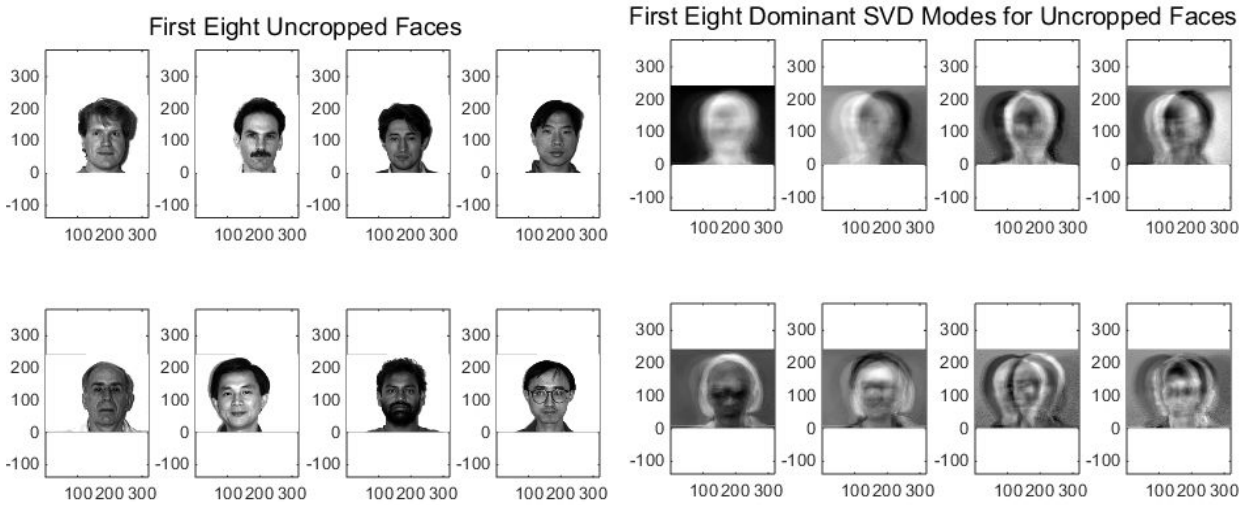

Figure 4: (Left) First eight different faces in uncropped Yale dataset. (Right) The eight dominant principle modes for uncropped faces.

The images on the left side of figure 4 are taken from the uncropped dataset, and the right side represent the principal modes generated from this set. The modes are quite uninformative in terms of facial structure and identifiability due to the lack of uniformity in face positioning in the original dataset. This highlights the importance of a standard format for the data samples when performing a principal component analysis.

Part 2

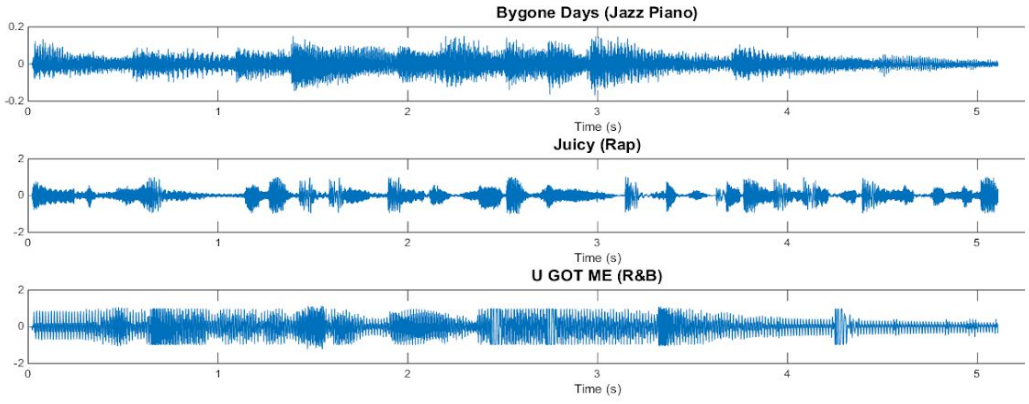

Figure 5: Sample five second clip from three different genres/artists.

Figure 5 demonstrates three examples of a five second clip from three different genres. In the matrix that the SVD is performed on however, the FFTs of these signals will be contained.

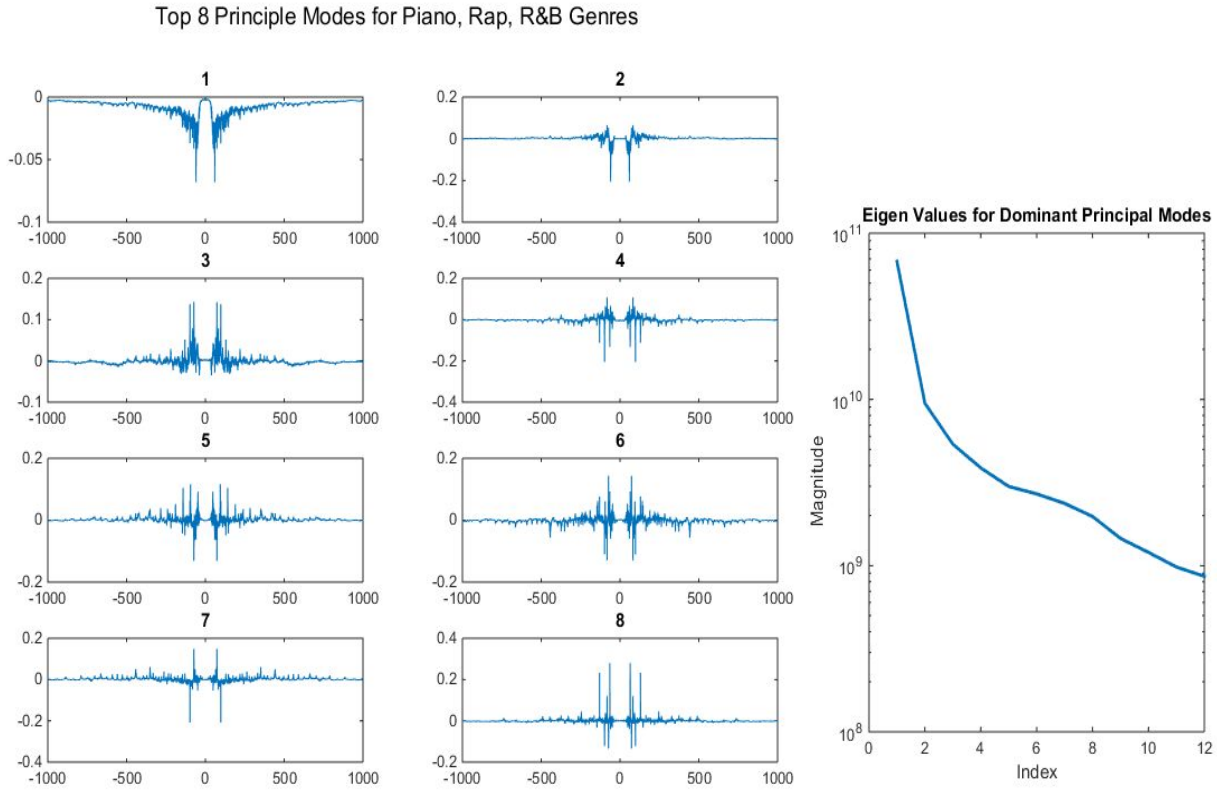

Figure 6: (Left) First eight dominant modes for three genres. (Right) The associated Eigenvalues for the principal modes.

The left side of figure 6 shows the 8 most important principal modes contained in the aforementioned matrix. Since these modes are generated from the frequency spectrums of the samples, the units of the x axis are Hz. On the right of the figure, the associated Eigenvalues are shown.

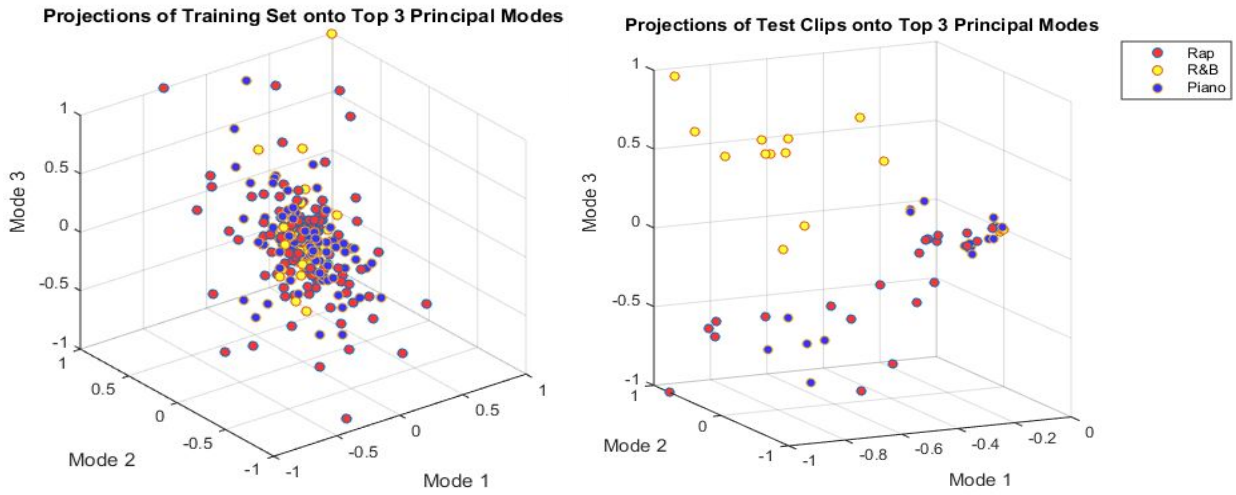

Figure 7: (Left) Projection of training set and (right) projection of test set onto three top principle modes for three different genres.

In an attempt to find three modes with the ability to differentiate between the three genres in question, I plotted each sample onto a three dimensional space where each dimension corresponds to one of the principal modes. The three colors correspond to different genres as indicated in the legend on the top right of figure 7. After trying many combinations of modes, it seemed that the genres did not possess sufficient separation in any of the principal mode-spaces. For this reason, I chose the three top principal modes to create a classifier. The right side of the figure demonstrates the projections of the test sets (about twenty each) onto these three modes. Interestingly, these test sets exhibit somewhat significant separation, especially between R&B (yellow) and the other two genres, though this may be due to chance.

V. Summary and Conclusions

In conclusion, the SVD has the potential to create extremely powerful and useful algorithms, such as classifiers for faces and genres, but it depends greatly on the training set, and which modes are selected for classification. In part 1, it was relatively easy to show that incorporating more and more modes into reconstruction created more detailed faces. However, in part 2, it was made clear that principal modes with significantly different profiles depending on the genre are not easy to find. This would require a much larger set of cleaner data with more different or *separable* genres or artists.

Appendix A MATLAB functions used

svd(X, 'econ'): The svd command computes u, s, and v, and the 'econ' option only produces the first m columns of v if m<n, and s is size [m x m].
fft(X): Generates the fourier transform of the column data of the matrix X.
diag(lambda): Outputs the main diagonal of a matrix in column form.
flipud(frame): flips the input matrix in the up-down direction.
semilogy(lambda): plots the input vector with the y axis in log format.