Math 444: Project 1

Levent Batakci

February 21, 2021

The work in this project focuses on data visualization and basic clustering. All of the mentioned implementations are coded in MATLAB and can be found in my public GitHub Repository.

Problem 1 - Model Reduction Data

The first data set consisted of data vectors $x^{(j)} \in \mathbb{R}^6$, with $1 \leq j \leq 4000$. The data is encoded as a matrix $X \in \mathbb{R}^{6\times 4000}$ and is loaded from the file *ModelReductionData.mat*. As an assessment of the raw data, we produced 2D scatter plots (Figure 1) corresponding to each pairing of components.

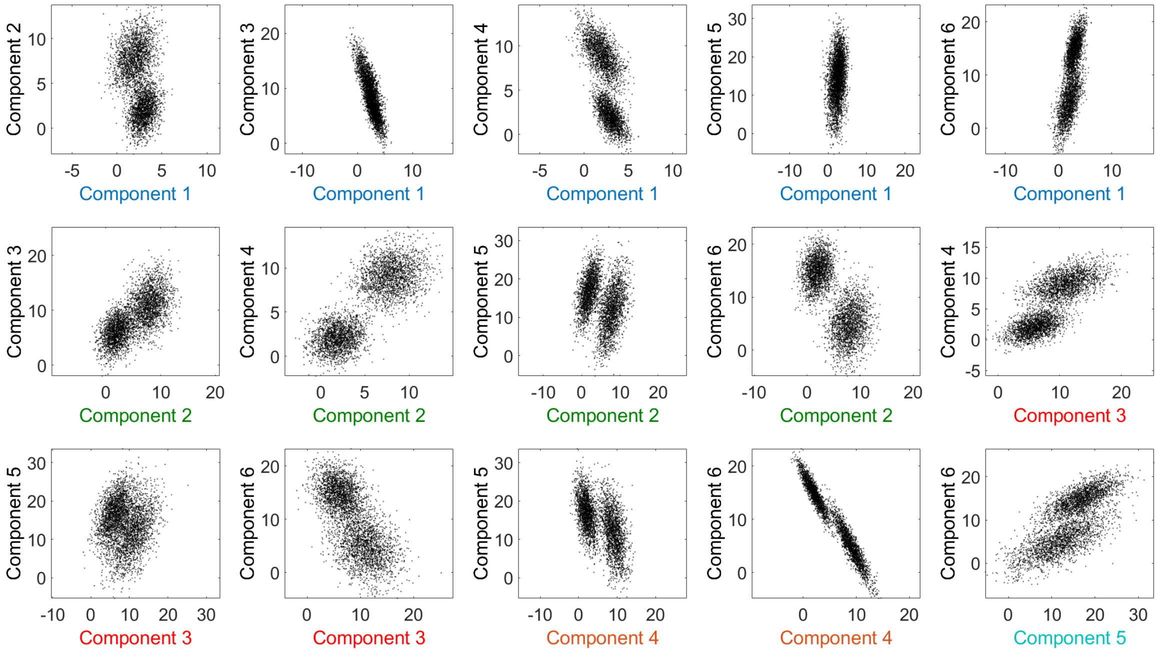

Figure 1: Raw Data Components Pair-wise Comparisons

No more than two apparent clusters are visible in any of the pair-wise plots. To simplify the problem of identifying clusters, we investigated whether the data's dimensionality could be reduced.

We decided to center the data so that performing a Principal Component Analysis would yield interpretations more accurate to how the data is distributed about its own center. When detecting clusters, this is generally preferable to working with interpretations about how the data distributes about the center of space (uncentered data). To compute the centered data X_c, we subtract the center of mass

$\bar{x} = \frac{1}{4000}\sum_{j=1}^{4000} x^{(j)}$ from each column of X. This essentially produces a shifted version of the data whose center is the origin.

After that, we compute the Singular Value Decomposition (SVD) of X_c. The SVD of X_c rewrites it as a product of a orthogonal matrix U, diagonal matrix D, and orthogonal matrix V^T (in that order). We computed the SVD of the centered data by using MATLAB's built-in algorithm 'svd'.

The columns of U can be viewed as feature vectors - they form a basis and reflect how the data distribute. The diagonal entries of D are the singular values of X_c and form a non-increasing, non-negative sequence. In a rough sense, the singular values reflect the relative importance of the associated feature vector in representing the data.

To be precise, larger singular values indicate a larger spread of the data when projected onto the associated feature vector. This is the key idea that drives Principal Component Analysis - comparing the scalar projections (principal components) of data onto feature vectors. The matrix of the first k principal components can be computed as $Z_k = [u_1 \cdots u_k]^T X$.

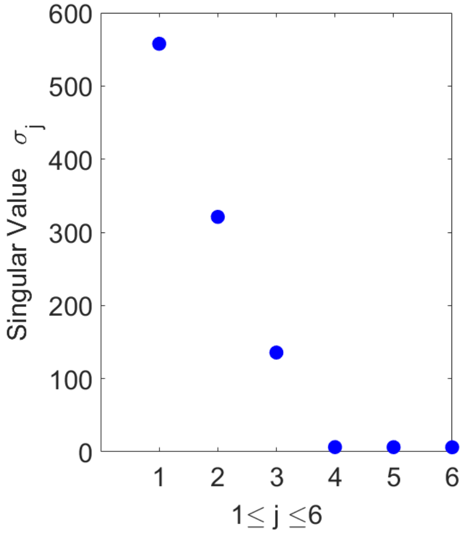

Figure 2: Centered Data (X_C) Singular Values

Once the SVD was computed, the singular values were extracted and plotted. The resulting plot (Figure 2) shows that the first three singular values are significantly larger than the rest. This leads to the conclusion that the data X effectively has a 3 dimensions.

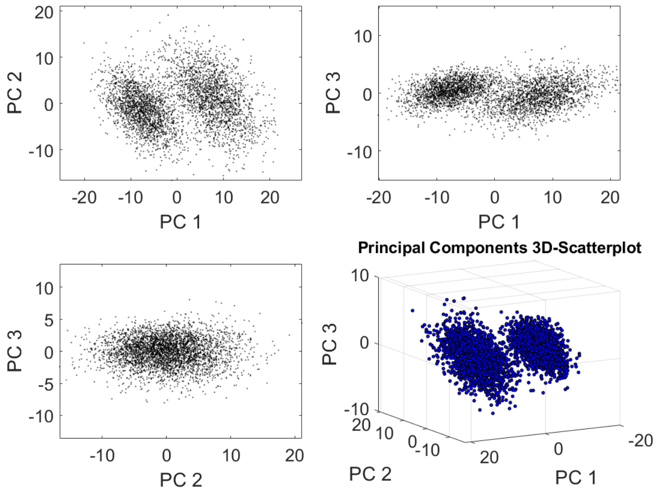

Figure 3: Centered Data (X_c) Principal Components (PC) Compared

From there, we computed the matrix Z_3 of the first 3 principal components. After that, we compared the first three principal components with 2D scatter plots and a 3D scatter plot (Figure 3). From these graphics, we were able to confirm that there are two apparent clusters in the data.

These results show practically the power of Principal Component Analysis. By reducing the dimensions of this data from 6 to 3, we were able to visually analyze the data and how its distributed. This enabled us to confidently conclude that the data consists of two main clusters.

Problem 2 - Biopsy Data

Problem 3 - Iris Data

Problem 4 - Aligned Face Data