Assignment 1

Duttaabhinivesh Devathi

March 20, 2017

Problem 1

This problem asks to visualize raw data in $\mathbb{R}^{6\times 4000}$ and project it in two dimensions. Because there are 6 dimensions in the data, there are $\binom{6}{2} = 15$ different combinations.
However, only a few projections are necessary to visualize the important aspects of the data that we are looking at. Seeing the data in the following projections shows that there are 2 clusters in the data.

The data can be found in `ModelReductionData.mat` and the the plots were created in `problem1.m`.

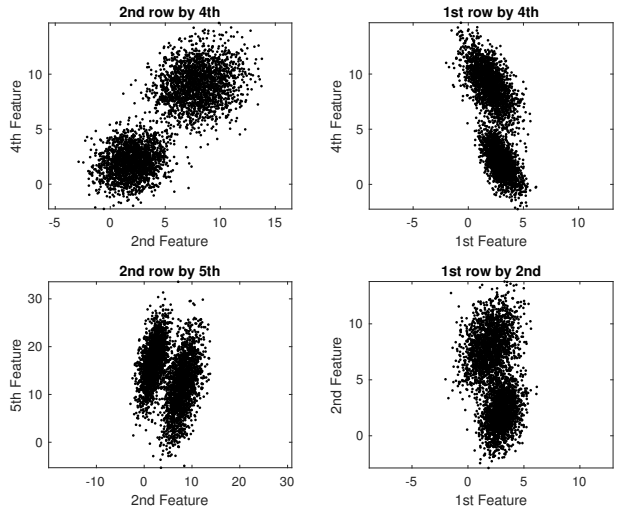

Figure 1: The figure shows different projections of the 6-dimensional data onto 2-dimensional space. The specific dimensions were chosen to highlight the two clusters. A fourth projection was included to show that the first three suffice and all other projections do not contribute important information about the data set.

As shown in Figure 1, there are clearly two clusters of data.

Figure 2 shows that there are a total of 6 singular values, but that the first 3 are two order of magnitude more dominant than the last 3. This means that while the data is still 6-dimensional, it can be approximated to 3 dimensions without much loss in the data. The right panel of figure two then plots $\binom{3}{2} = 3$ combinations of the first three principal components of the centered data, confirming that there are two clusters.

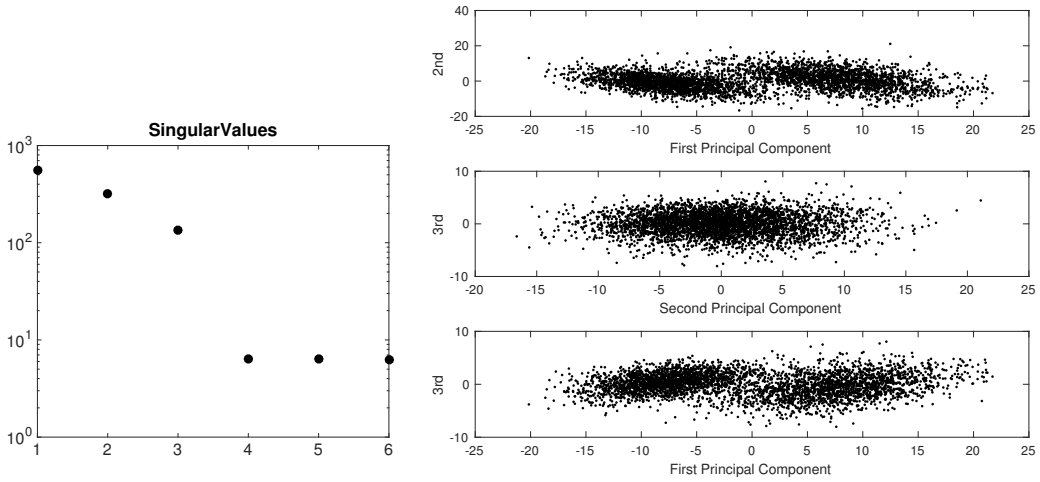

Figure 2: The left figure plots the singular values of the centered data according to magnitude. The right figure shows plots of the combinations of the first three significant principal components. The top is the 1st and 2nd, middle is 1st and 3rd, and bottom is 2nd and 3rd.

Problem 2

This problem is an exercise in approximating handwritten digits with the principal components of the data of each digit. Each digit was approximated with different numbers of principal components, k = 5,10,15,20,25. The data is found in `HandWrittenDigits.mat`

The matlab code below shows the construction of each of the feature vectors for a specific digit, and then creates an approximation with each k (each row) with 5 different samples (each column). It can be found in the file `problem2_resubmit.m`.

```
Il = I == nums(l);
Xl = X(:,Il);
[U,~,~] = svd(Xl);
figure(l)
for i = 1:5
Uk = U(:,1:5*i); Zk = Uk'*Xl;
    for j = 1:5
        xj = Uk*Zk(:,j);
        residuals(i,j) = norm(Xl(:,j) - xj,2);
        subplot(5,5,(i-1)*5 + j)
        imagesc(reshape(xj,16,16)');
        colormap(1-gray);
        axis('square')
        axis('off');
    end
end
```

The code takes the Singular-Value Decomposition of the entire set of datapoints for each digit, then makes a rank k approximation, where k ranges from 5 to 25 by 5. Then, the first five feature vectors from each approximation are plotted using a grayscale map. `Residuals` represents the norm of the difference between the actual digit and the approximated for each digit.

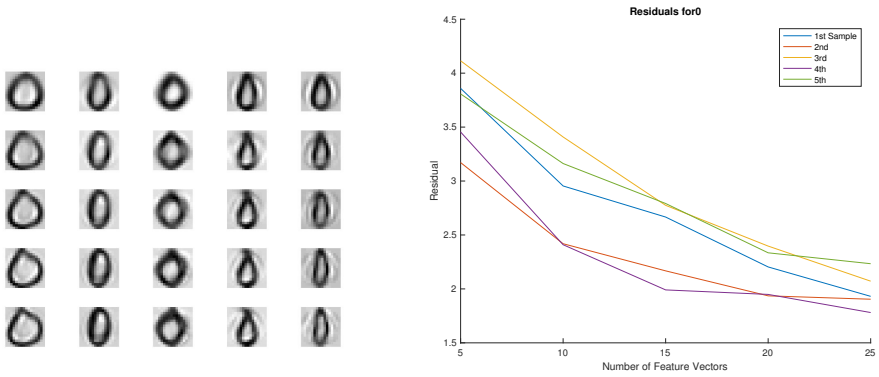

Figure 3: The left panel shows different approximations for handwritten 0s. Along the row are different samples of 0s. And along the column are closer approximations by number of principal components (k = 5,10,15,20,25). The right panel is a plot of the residual of each approximated digit against the number of feature vectors of the approximated date.

Figure 3 shows approximations for different samples of handwritten 0's. Only the 0's are shown here because the plots for the other digits 1, 4, and 7 are similar and would be redundant. The left panel of

Figure 3 clearly shows that the residual between the approximated digit and the actual digit decreases as you increase k, the number of feature vectors in the data.

Problem 3

This problem is very similar to the first one, and asks us to identify three different clusters of flowers (corresponding to each species of flower). The 2-d projections of each of the 4 dimensional criteria are shown below. The data can be found in `IrisData.mat`.

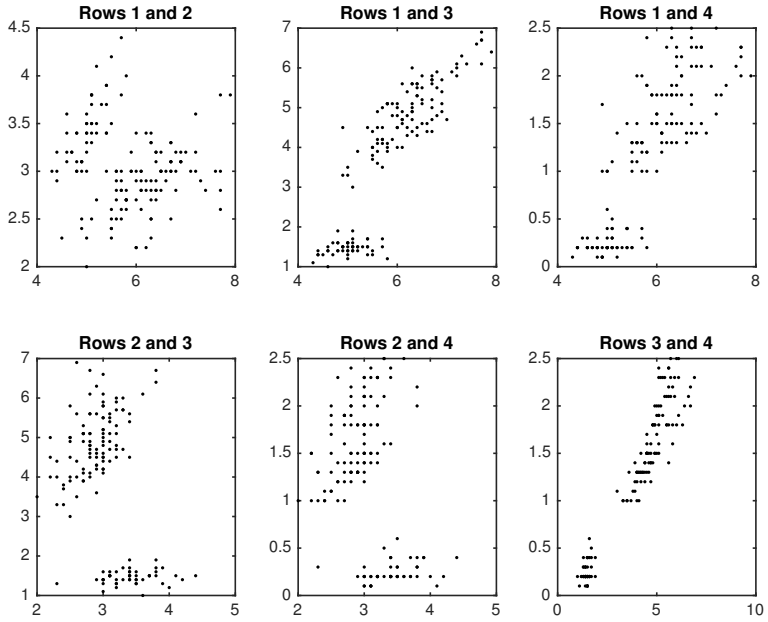

Figure 4: This shows all the possible 2-dimensional projections of each of the features of the flowers.

Although there are two visible clusters (one large, and one small), the large corresponds to two different species of flowers, which have very similar characteristics. A Principal Component Analysis will show the significance of each feature vector in the flowers.

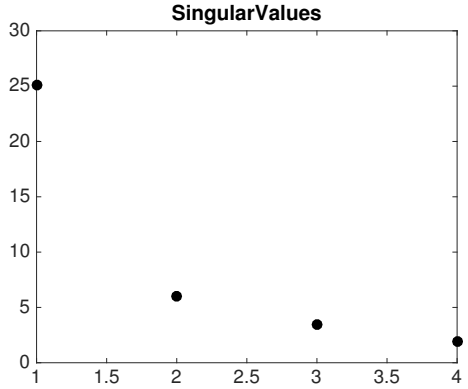

Figure 5: The singular values of the Iris data are plotted in order of the diagonal.

There is clearly one principal component that defines the flower, and the other three have a high correlation in each species of flower. Further clustering analysis must be done in order to properly identify each flower.