Assignment 1

Duttaabhinivesh Devathi

February 9, 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.

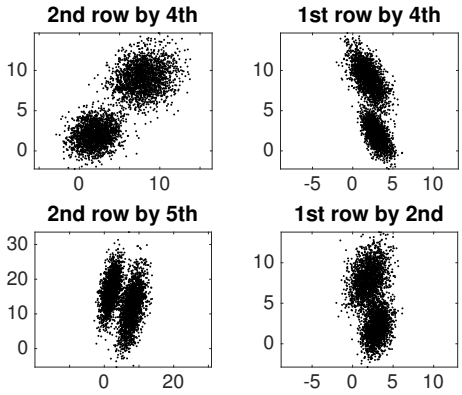
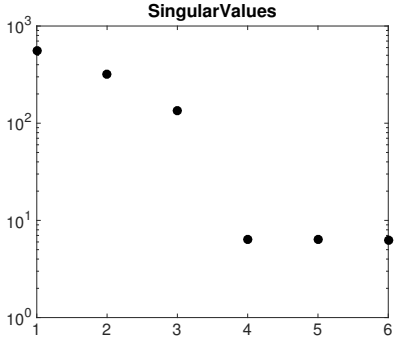


Figure 1: The figure shows different projections of the 6-dimensional data onto 2-dimensional space. The specific dimensions were chosen to highlight the two clusters. A fourth projection was included to show that the first three suffice and all other projections do not contribute important information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

As shown above, there are clearly two clusters of data.



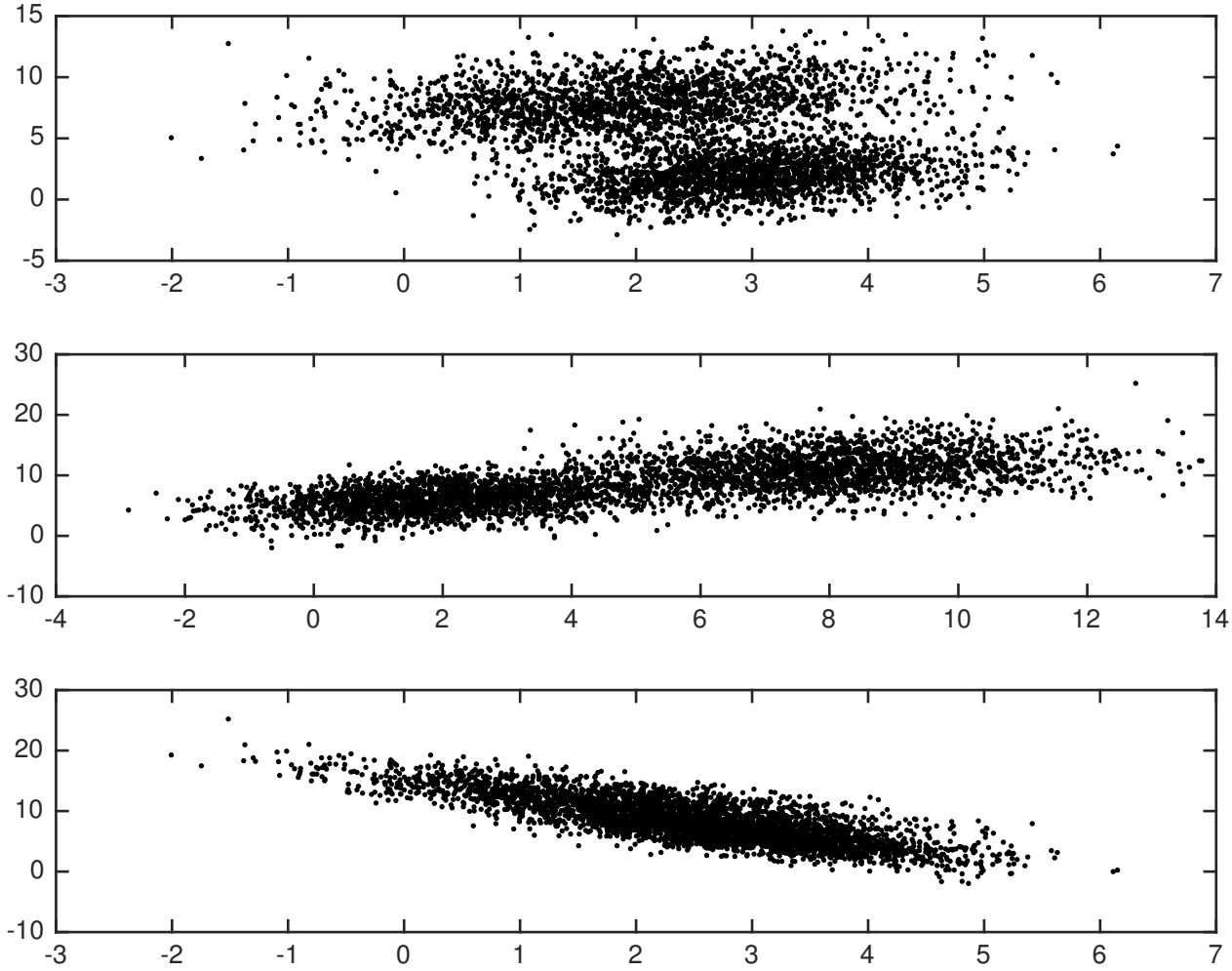


Figure 2: Each of these plots is a combination of the first few principal components. The top is the 1st and 2nd, middle is 1st and 3rd, and bottom is 2nd and 3rd.

Below are the plots of the first 3 principal components.

With each of these plots (except maybe for the 3rd), there is a clear indication of two clusters in the data.

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 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).

```
figure(1)
I_i = find(I == k1);
X_i = X(:,I_i);
[U,D,V] = svd(X_i);
residuals = zeros(5,5);

for i = 1:5
    Ur = U; Dr = D(1:5*i,1:5*i); Vr = V(1:5*i,:);
    Z = Dr*Vr;
    for j = 1:5
        x_j = 0;
        for k = 1:(5*i)
            x_j = x_j+Z(k,j)*U(:,k);
        end
        subplot(5,5,(i-1)*5 + j)
        imagesc(reshape(x_j,16,16)')
        colormap(gray);
        axis('square')
        axis('off');
        residuals(i,j) = norm(X_i(:,j) - x_j);
    end
end
```

Below is a figure of approximations for different samples of handwritten 0s.

To understand the effectiveness of adding more principal components, the norm of the difference of the actual feature vector and the approximation must be calculated and graphed against the number of feature vectors k. This looks like below:

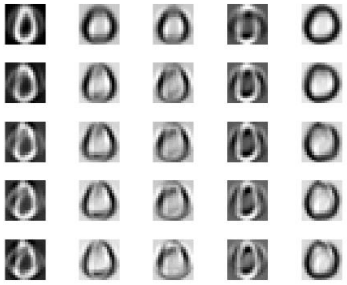
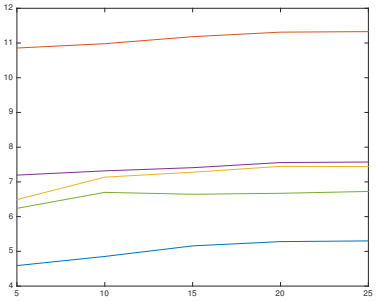


Figure 3: Shows different approximations for handwritten 0s. Along the columns are different samples of 0s. And along the rows are closer approximations by number of principal components (k = 5,10,15,20,25).



The issue with this plot is that the residual plot did not perform properly, and this is as in issue with the matlab code that has not been resolved. However, the desired result is that the norm of the error decreases concave up as you add more feature vectors.

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 2d projections of each of the 4 dimensional criteria are shown below.

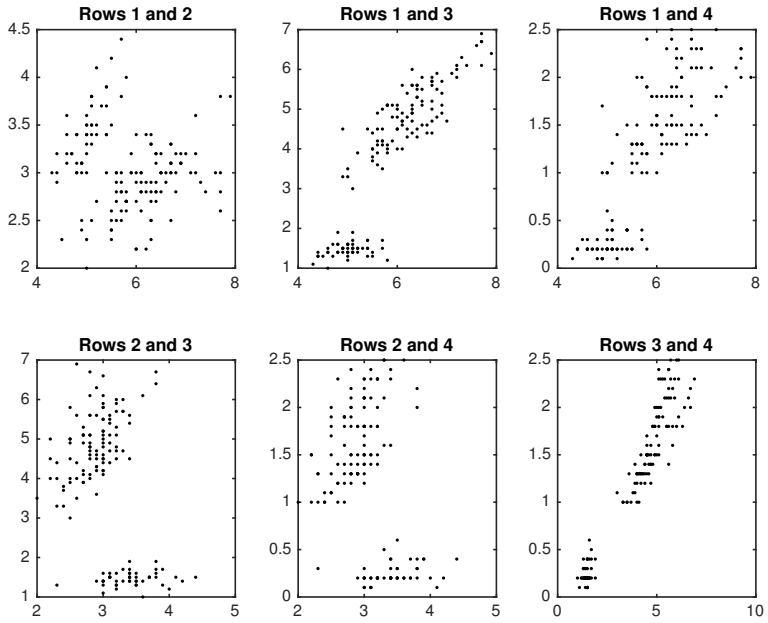


Figure 4: This shows all the possible 2-dimensional projections of each of the features of the flowers.

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 if there are more than 2 feature vectors.

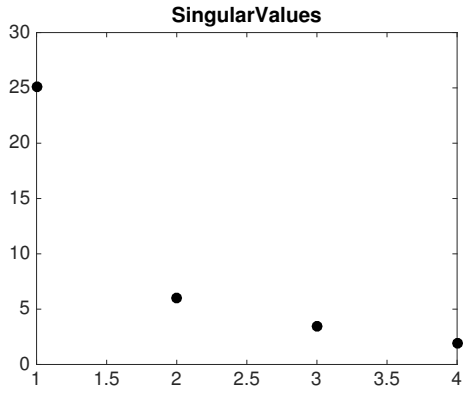


Figure 5: This shows all the possible 2-dimensional projections of each of the features of the flowers.

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.