Math 444: Homework 1

Iris Zhang

February 4, 2020

Problem 1

To visualize the raw data in three dimensions, I plot the 3D graph in Matlab, see Figure 1.

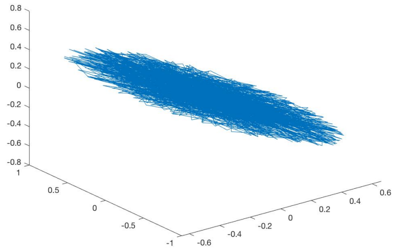

Figure 1: 3D graph of the raw data

From the Figure 1 above, we can see that the data roughly lies in a plane.
Then, we select two components of the data at one time, see Figure 2. We can see one cluster in each graph.

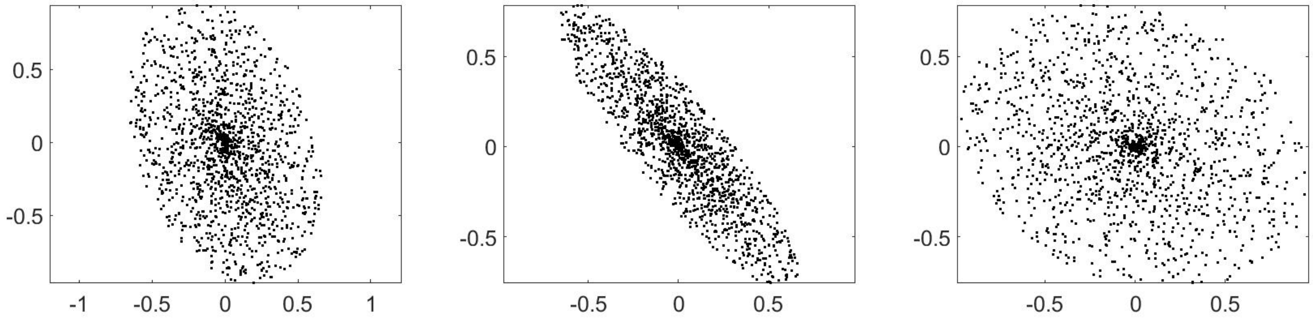

Figure 2: scatter plots of the raw data (Left: x_1 and x_2 component; Middle: x_1 and x_3 component; Right: x_2 and x_3 component)

To center the data and compute the SVD, I include a snippet of my Matlab code:

```
% center the data
xavg = 1/1500*sum(X,2);
Xc = X-xavg*ones(1,1500);
% compute SVD for the centered data
[U,D,V] = svd(Xc);
singularvalue = diag(D);
```

The singular values are shown in Figure 3. We can clearly see that the first two values are nonzero, which implies that the effective dimensionality of the data is 2.

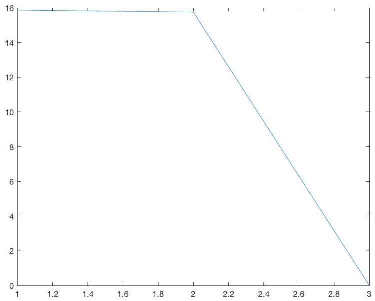

Figure 3: singular values for centered data

After calculating the principal components, the three scatter plots are shown in Figure 4. Compared with three graphs in Figure 2, we can see that the data are more centered. Also, there is a presence of clusters in the graph of the first two dominant principal components.

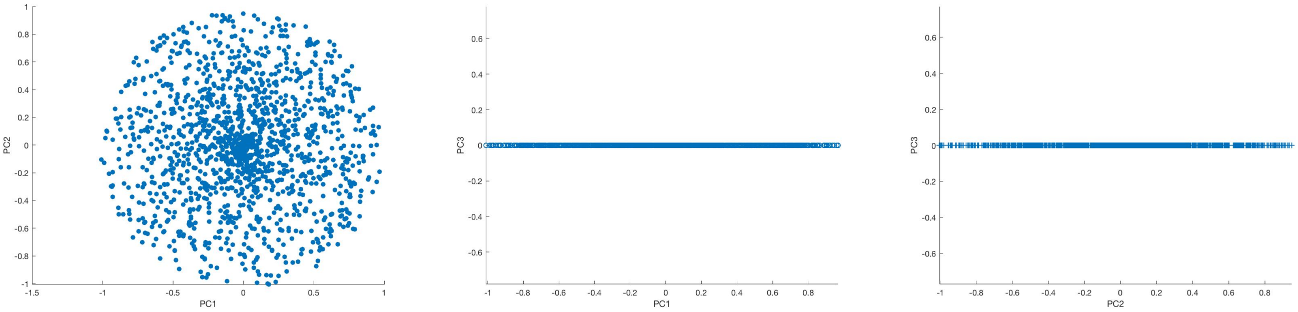

Figure 4: scatter plots for principal components

Problem 2

To visualize the data, we use the scatter plots, see Figure 5. From that graph, we can see clusters in each graph.

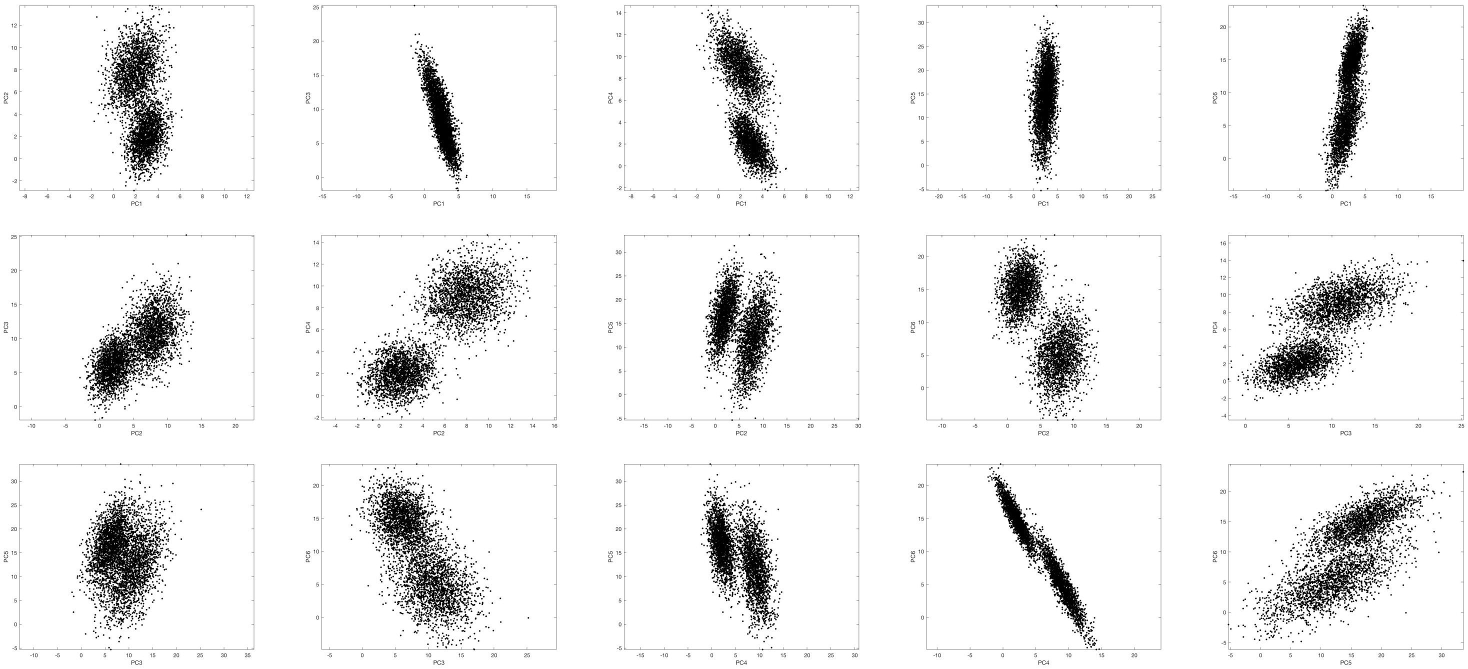

Figure 5: scatter plots, the coordinates are shown in the graphs

Now, we center the data and compute the SVD of the centered data. The plot of the singular values are shown in Figure 6.

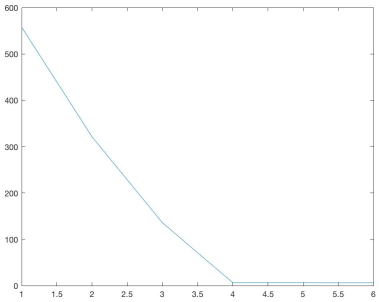

Figure 6: singular values for centered data

We can see from the graph that the first three singular values are largely different from zero while the last three are roughly zero, which means that the effective dimensionality of the data is 3.

Now we plot the scatter plots of the first three principal components, see Figure 7. From the graph, we can see that there are presences of clusters in the data.

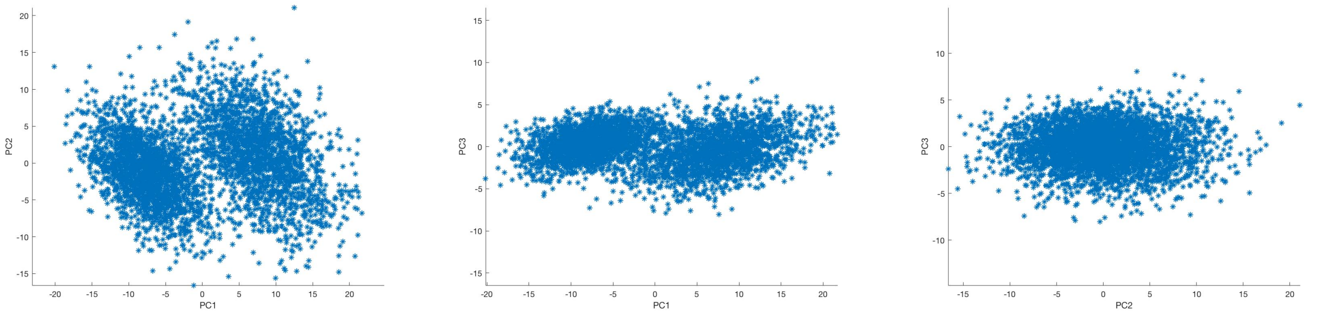

Figure 7: scatter plots of the three dominant principal components

Problem 3

We approximate the sample specimen by linear combination of the first k feature vectors. I include a snippet of my Matlab code.

```
value2 = find(I==2);
X2 = X(:,value2);
[U2,D2,V2] = svd(X2);
sigma2 = diag(D2);
for k=1:5
    subplot(4,5,k)
    x2j_app=0;
    for l=1:(5*k)
        x2j_app = x2j_app + sigma2(l)*V2(1,l)*U2(:,l);
    end
    figure(1)
    imagesc(reshape(x2j_app,16,16)')
    colormap(1-gray);
    axis('off')
    axis('equal')
end
```

The approximation images see Figure 8. We can see that with the increase of the number of principal components, the images become more and more clear.

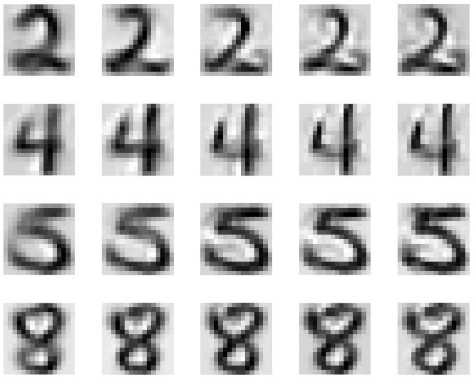

Figure 8: Approximation image left to right: k=5,10,15,20,25

The residual with $k = 25$ see Figure 9. I include a snippet of my Matlab code for plotting the residual with k=25.

```
subplot(2,2,1)
    x2_app = 0;
    for l =1:25
        x2_app=x2_app+sigma2(l)*V2(1,l)*U2(:,l);
    end
    x2_res=X2-x2_app;
 imagesc(reshape(x2_res(:,1),16,16)')
    colormap(1-gray);
    axis('off')
    axis('equal')
```

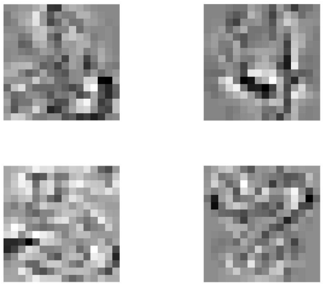

Figure 9: residual with k=25

The norms of the errors as a function of k see Figure 10.

I include a snippet of my Matlab Code:

```
for k = 1:30
    Uk = U(:,1:k);
    Zk = Uk'*Xk;
    xj = Uk*Zk(:,1);
    residuals(k,1) = norm(Xk(:,1) - xj,2);
end
figure(10)
subplot(2,2,1)
plot(1:1:30,residuals(:,1))
title(strcat('norms of error for  ',' ', digit))
```

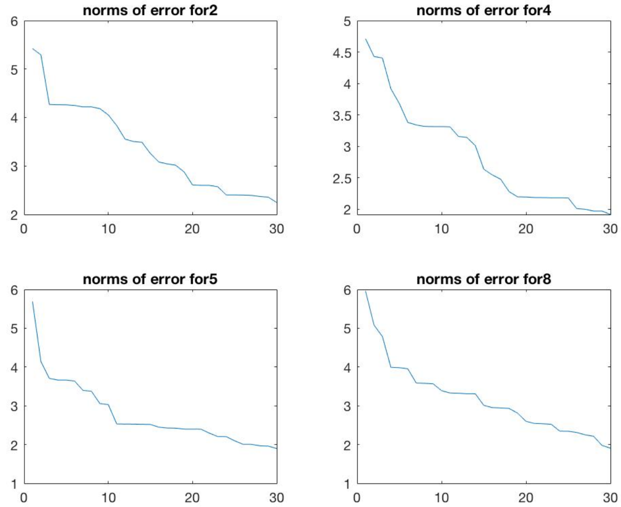

Figure 10: Norms of error

Problem 4

First, we look at the scatter plots of the raw data, see Figure 11. We can roughly see two clusters in the plots, neglecting the color of each species.

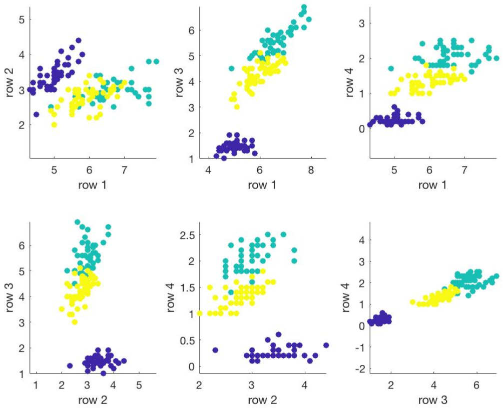

Figure 11: scatter plots of the raw data

We then do the PCA (with centering the data) using Matlab and plot the singular values, see Figure 12. I include a snippet of my Matlab code.

```
% center the data
n = size(X4,2);
Xmean = mean(X4,2) ;
A = X4 - Xmean*ones(1,n);

% SVD
[U,S,V] = svd(A,'econ');
Z = S(1:3,1:3)*V(:,1:3)';

% plot sigma
sigmairis = diag(S)
figure(5)
plot(sigmairis)

figure(1);
scatter(Z(1,:),Z(2,:),17,I,'filled')
xlabel('PC1'); ylabel('PC2')
figure(2);
scatter(Z(1,:),Z(3,:),17,I,'filled')
xlabel('PC1'); ylabel('PC3')
figure(3);
```

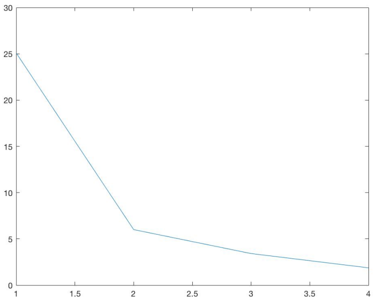

Figure 12: singular values

We can see that the first two singular values are far away from zero. And also, the first one should be dominant. And the effective dimensionality of this data set is two. Then we make the scatter plot of the principal components, see Figure 13. For completeness, I also include the scatter plot containing the third principal componentsin Figure 13.

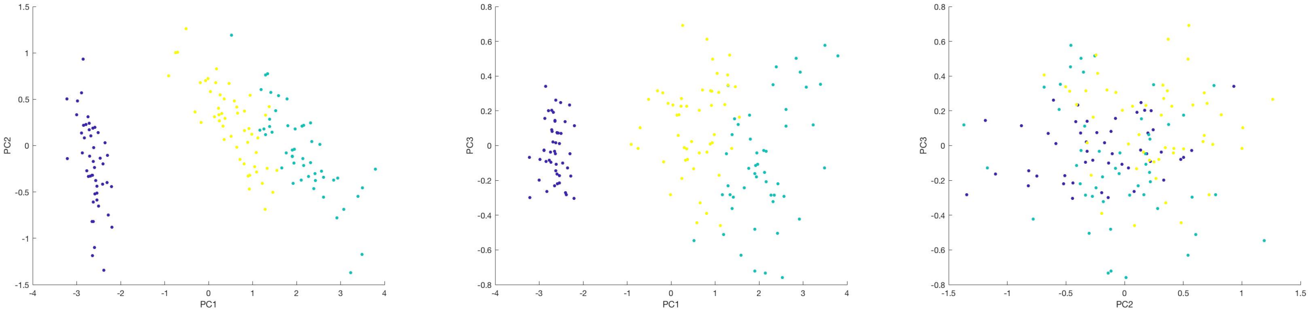

Figure 13: scatter plots of the principal components

In the first two graphs of Figure 13, we can see that there are clusters for each different species, one is clearly separated from the other two. However, the yellow and green ones do not have a large distance from each other. To separate them, we can do the clustering!