# Math 444: Homework 3

Iris Zhang

February 28, 2020

## Problem 1

The Matlab code for LDA is attached in the appendix I.

## Problem 2

For **WisconsinBreastCancerData**, we first compute the LDA separation direction and we plot the projections of the two different clusters on the separators, see Figure 1.

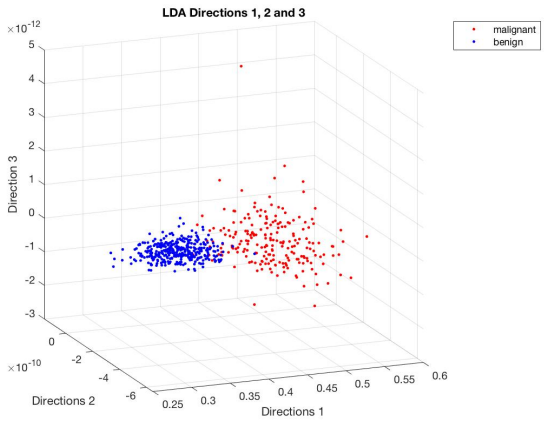


Figure 1: projection of clusters on separators

We can see that the graph above have two relatively separated clusters.
We then plot the histogram of the projections of the two classes, see Figure 2. We can see that the clusters are not completely separated in Direction 2 and 3 while in Direction 1, two clusters are roughly separated with small proportion of overlap.

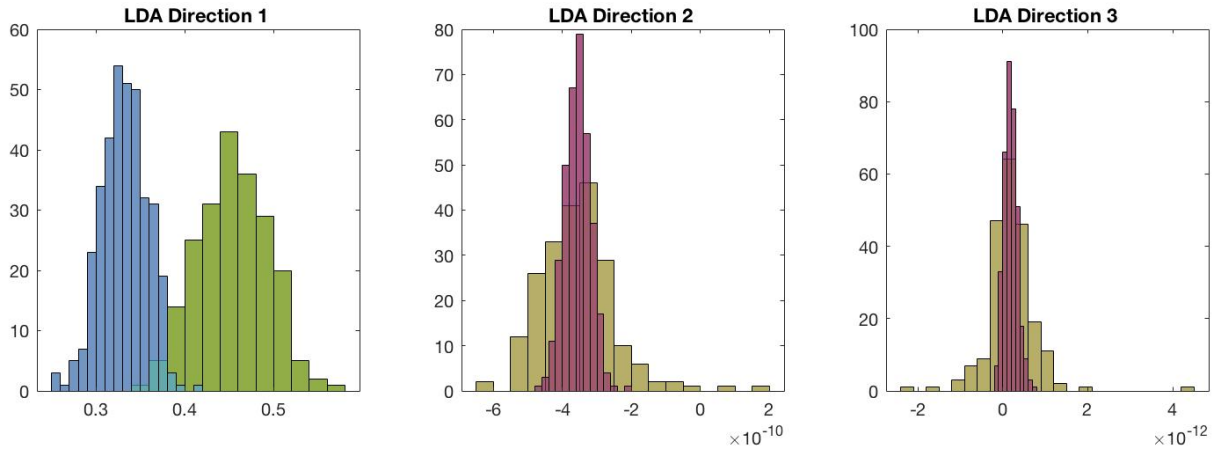


Figure 2: histogram of projections of two classes

Then, we do PCA to reduce the dimentionality and plot the singular values, see Figure 3.

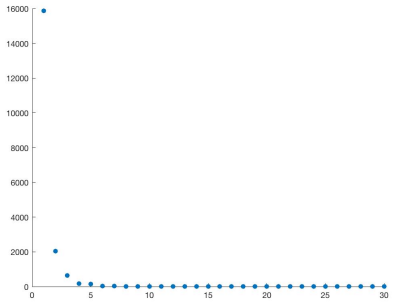

Figure 3: singular values

We can see that the first singular value dominant. For completeness, we continue to do the plot of the first three principal components, see Figure 4.

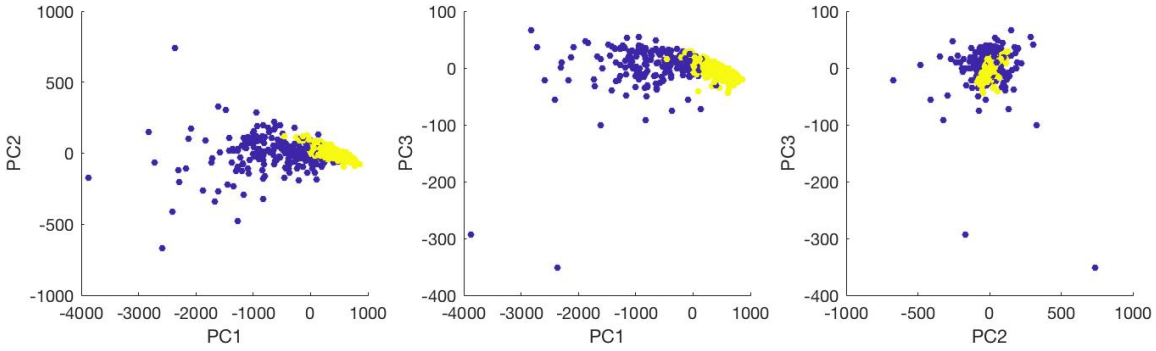


Figure 4: principal components

Then, we do LDA on the reduced data matrix and plot the projections of the two different clusters on the separators, see Figure 5.

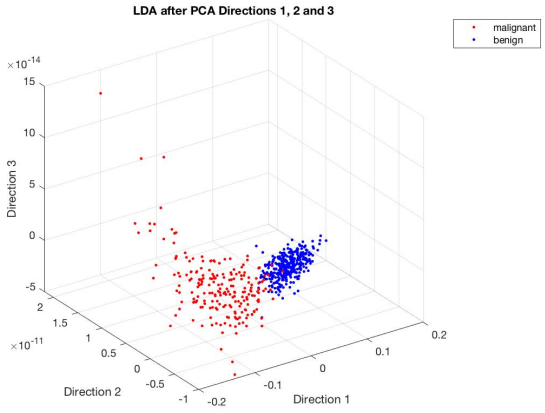


Figure 5: projection of clusters on separators

We then plot the histogram of the projections of the two classes, see Figure 6. We can see the clusters are still not completely separated.

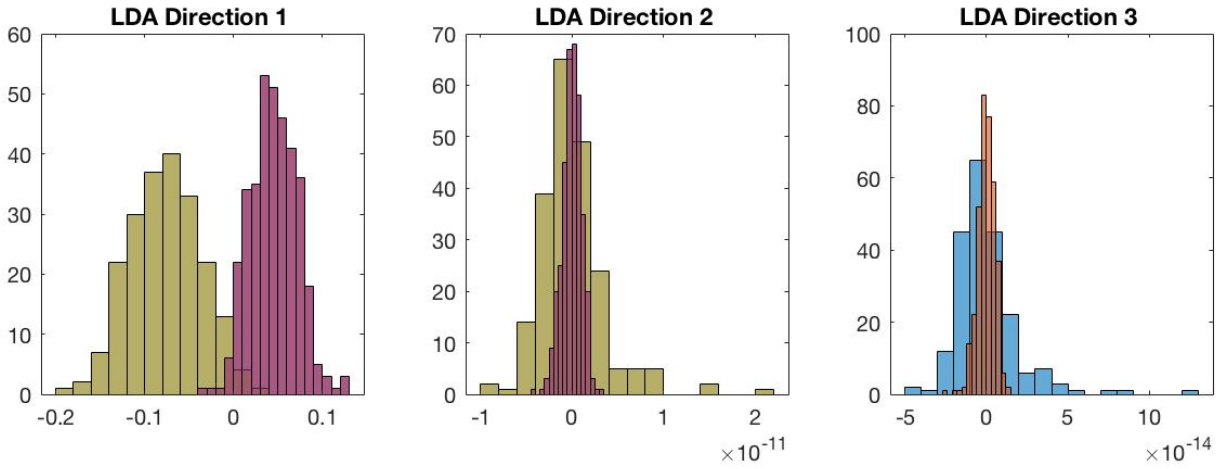


Figure 6: histogram of projections of two classes

Besides, we try to use the first ten largest components and then apply LDA to see the separation. Also, we plot the histogram for this, see Figure 7.

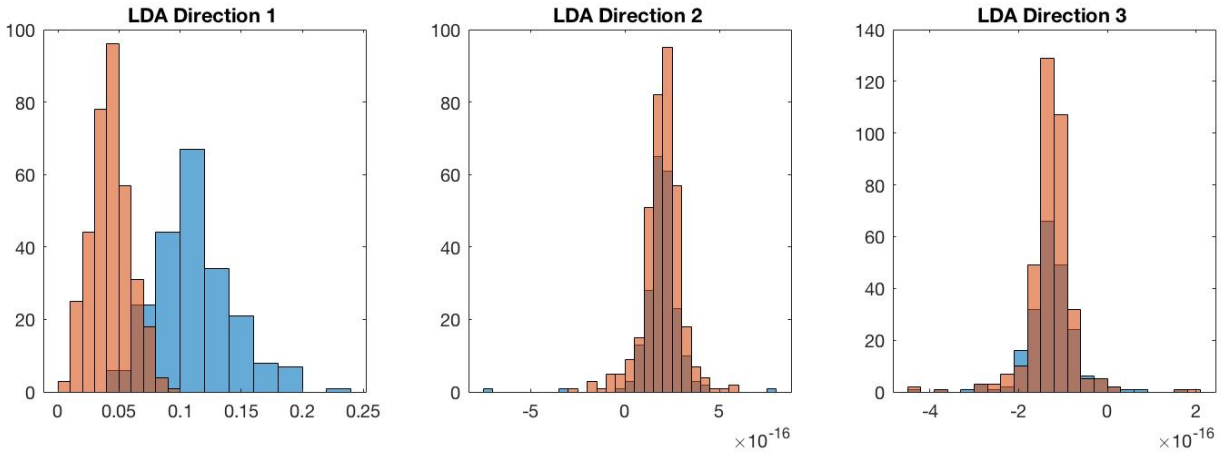


Figure 7: histogram of projections of two classes for first ten largest components

From above graphs, even we reduce the dimentionality by selecting only the first ten largest components, the clusters are not completely separated.
Overall, the clusters are not completely separated after we do PCA and LDA, even in the first direction, which implies that those attributes cannot well indicate whether the patient is malignant or benign.

# Problem 3

First, we plot the original data set, see Figure 8.

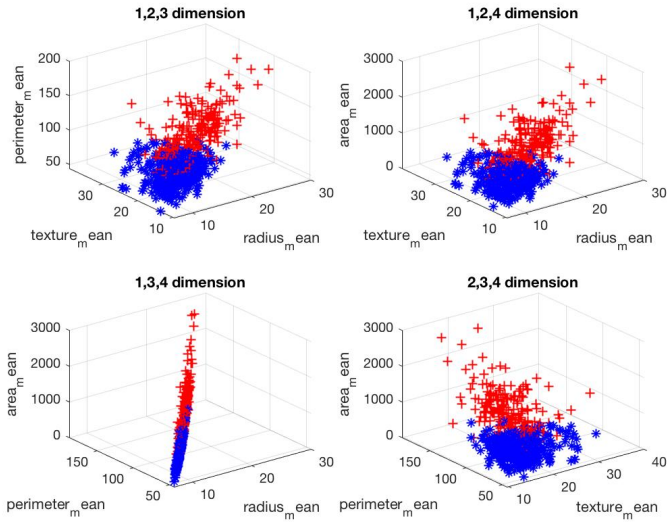


Figure 8: original data

Now we apply the k-means and k-medoids to **WisconsinBreastCancerData**, see Figure 9 and 10, respectively.

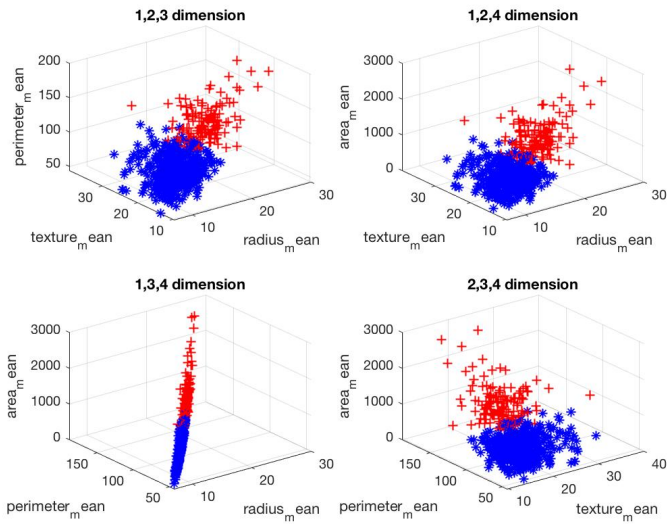


Figure 9: data after doing k-means

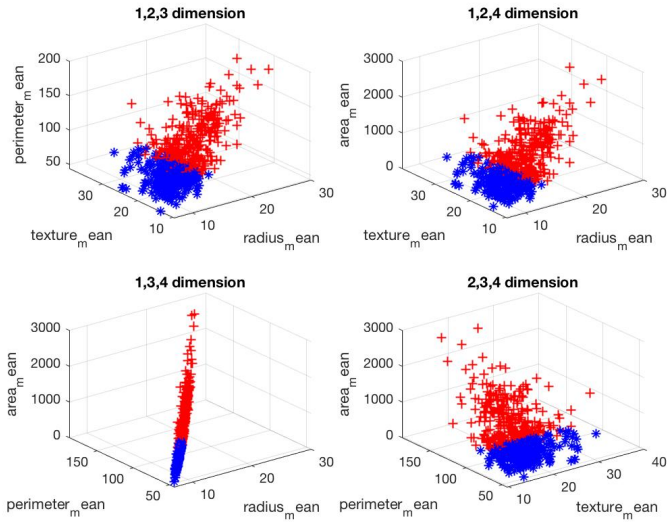


Figure 10: data after doing k-medioids

Compared to the original data clusters, we can find that the clusters generated by k-means algorithm is closer to the original separation.
Besides, in order to determine how much do they agree with the original labels of the data, we use the matrix $C_{2X2}$ and $D_{2X2}$. As a result, we can find that the ratio is always 212:357. I also include a snippet of my Matlab code.

```
D = zeros(2,2);
D_11 = 0;
D_12 = 0;
D_21 = 0;
D_22 = 0;
```

```
for i = 1:size(A3,2)
    if (I3c(1,i) == 1 & ismember(size(A3,2) , malignant))
        D_11 = D_11+1;
    end
    if (I3c(1,i) == 1 & ismember(size(A3,2) , benign))
        D_12 = D_12+1;
    end
    if (I3c(1,i) == 2 & ismember(size(A3,2) , malignant))
        D_21 = D_21+1;
    end
    if (I3c(1,i) == 2 & ismember(size(A3,2) , benign))
        D_22 = D_22+1;
    end
end
D = [D_11 D_12; D_21 D_22]
```

# Problem 4

We compute the LDA separator based on k-means classifications, and plot the histograms of the two classes, see Figure 11.

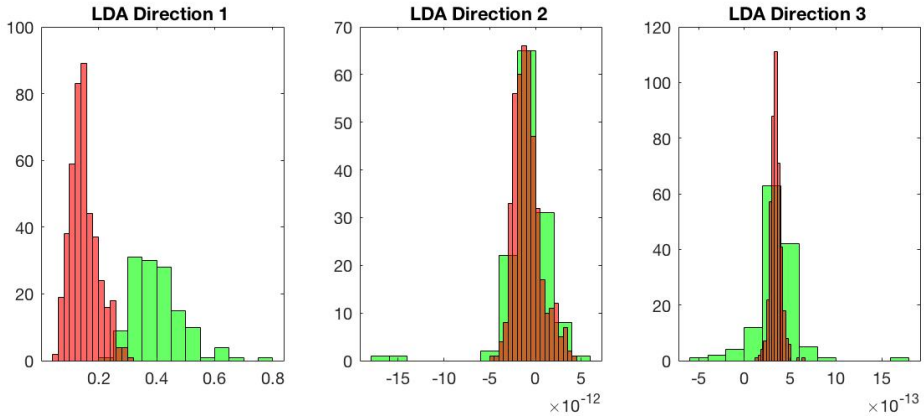


Figure 11: histogram of LDA separator after k-means

Then, we compute the LDA separator based on k-medioids classifications, and plot the histograms of the two classes, see Figure 12.

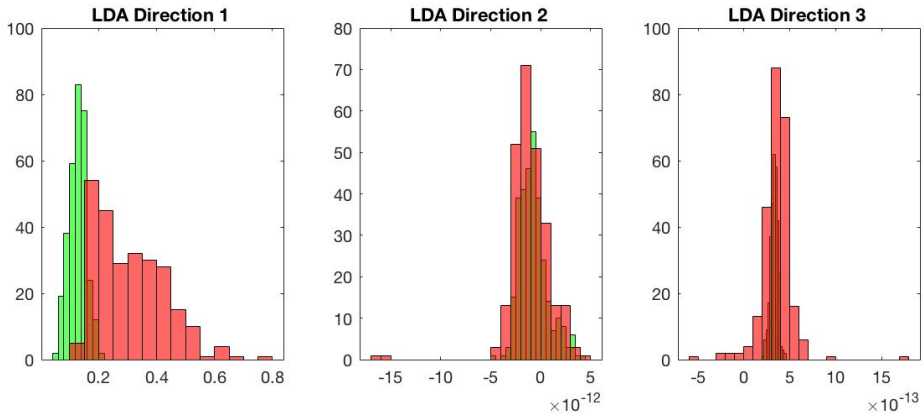


Figure 12: histogram of LDA separator after k-medioids

From above graphs, we can see that the clusters, green and red, are not completely separated, which indicates that those attributes cannot determine the condition of the patient well.

# Problem 5

For **WineData**, we first compute the first two LDA separation direction and we plot the projections of the three different clusters on the separators, see Figure 13.

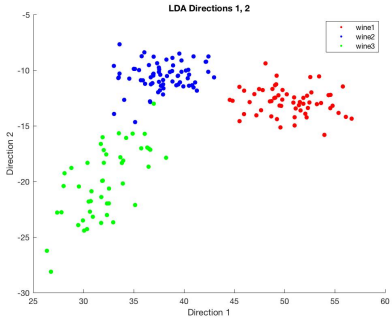


Figure 13: projections of the three different clusters on the separators

From the above graph, we can see clear separated data of three different types of wine, which is much clearer than what we did in homework 2. We also plot the histogram for it, see Figure 14.

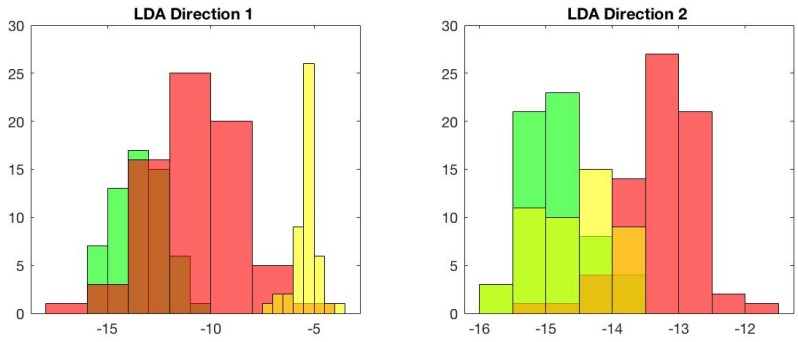


Figure 14: histogram of LDA separator

Then, we pick first five largest components and plot the histogram again, see Figure 15.

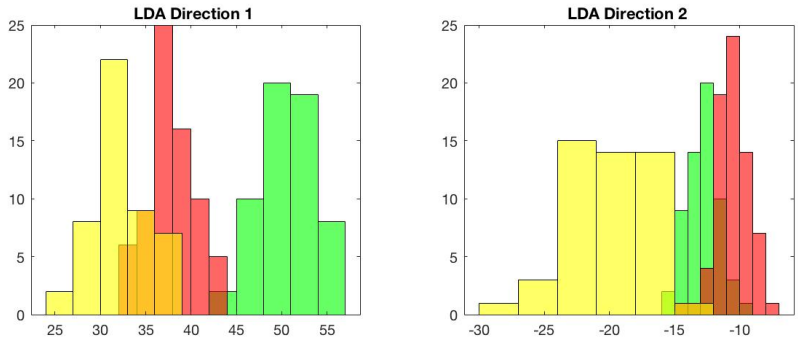


Figure 15: histogram of LDA separator for first ten largest components

From above two graphs, we can see that the clusters are more separated from each other after we reduce the dimensionality.

# Problem 6

For **ForestSpectra.mat**, we first compute PCA to determine its effective dimensionality. Then we plot the singular values in Figure 14.

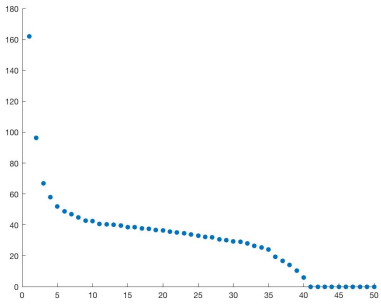

Figure 16: singular values

We can see that the first two singular values are dominant. Then, we compute the LDA separation direction and we plot the projections of the three different clusters on the separators, see Figure 15.

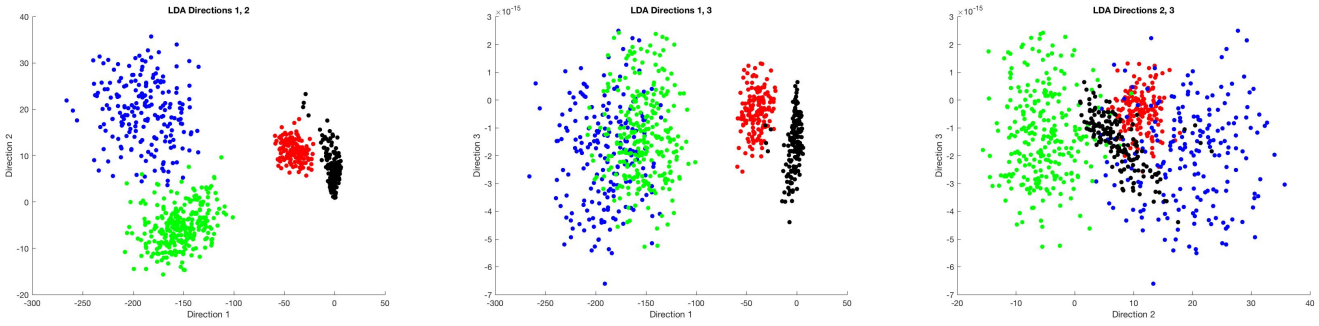


Figure 17: projection of clusters on separators

We can see relatively clear separated clusters in the first graph while there are overlaps between the clusters in the rest of two graphs.

## Appendix I : Matlab code for LDA

```
function [V,D,Xw] = LDA(X,I,clusters_num)
tau = 10^-16;
[n,p]=size(X);

% calculate the global mean
c = 1/p*sum(X,2);

Xw = [];
% calculate the cluster means
for i = 1:clusters_num
    X_{i} = X(:,I==i);
    b = size(X_{i},2);
    c_{i} = 1/(b)*sum(X_{i},2);
    Xc_{i} = X_{i} - repmat(c_{i},1,b);
    Xw = [Xw Xc_{i}];
end

Sw = Xw * Xw';
Sw = Sw + tau *eye(size(Sw));
Sb = zeros(n,n);

for i = 1:clusters_num
    b2= size(X_{i},2);
    Sb = Sb + b2*(c_{i}-c)*(c_{i}-c)';
end

Sb = Sb + tau*eye(size(Sb));
A = Sw\Sb;
[D,V] = eigs(A,3,'LM');

end
```