MATH444 HW3

Jiasen Zhang

1 Problem 1

Test my LDA algorithm in Figure 1 using Gaussian distribution data. The group centers are:

$$\begin{bmatrix} -2.9958 \\ -3.0944 \\ -3.1437 \end{bmatrix} \quad \begin{bmatrix} 2.0308 \\ 0.1135 \\ -0.0254 \end{bmatrix} \quad \begin{bmatrix} 4.0009 \\ 4.0621 \\ -6.0079 \end{bmatrix}$$

By computing the two LDA separating directions, the 3 clusters are classified well.

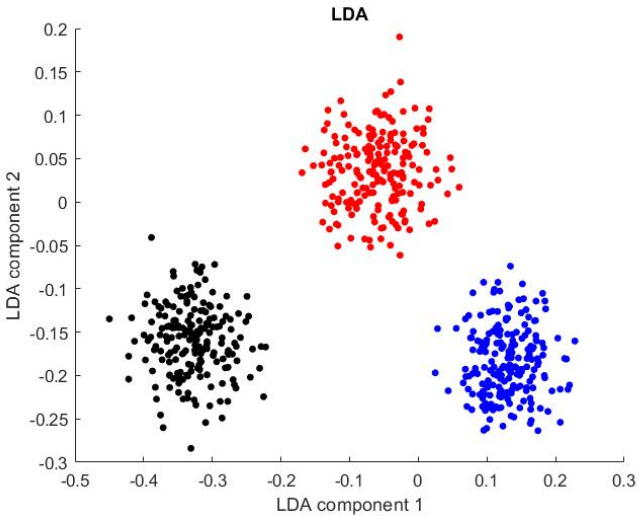

Figure 1: Problem 1

2 Problem 2

Figure 2 shows the plots of the two LDA directions(left) and the first two principal components(right).

Result of LDA is similar to that of PCA, the blue cluster is seperated well, while the black and red clusters are close to each other.

However, the LDA result is better. By projecting along the maximally seperated direction, the black and red clusters in LDA are more distinguishable. There are less points in the mixing area between the black and red clusters.

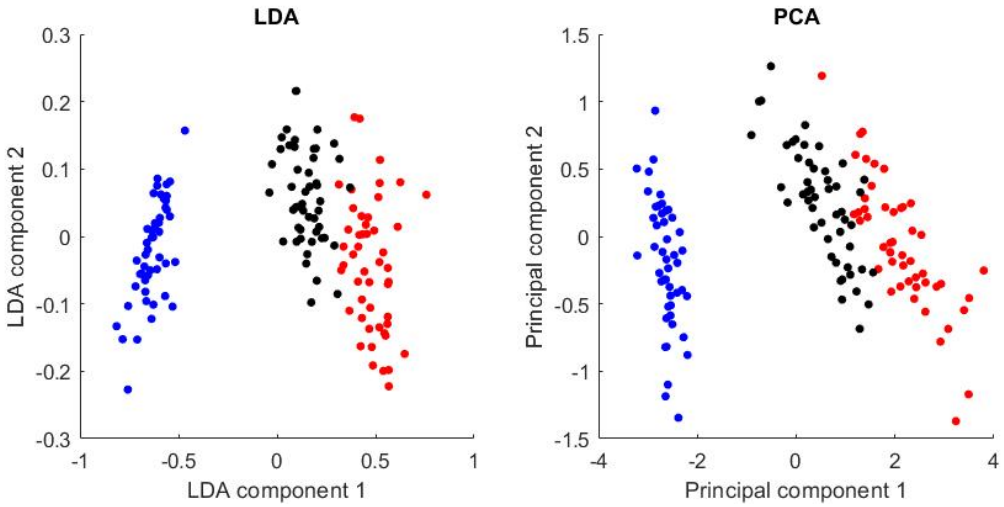

Figure 2: Problem 2: IrisData

3 Problem 3

3.1 (a)

The LDA result is shown in the left Figure 3. In my opinion, the attributes don't carry enough information to separate the two clusters, because there are still a small overlapping area between the two clusters. But I think the differences between the two clusters are characterized well, because the peaks of two clusters are not close to each other, and the overlapping area is small.

3.2 (b)

The PCA result is shown in the right Figure 3. Both of the two results have two peaks for each cluster, and have an overlapping area. However, the result of LDA is much better. For LDA result, the peaks of the two clusters are farther, and the overlapping area is smaller.

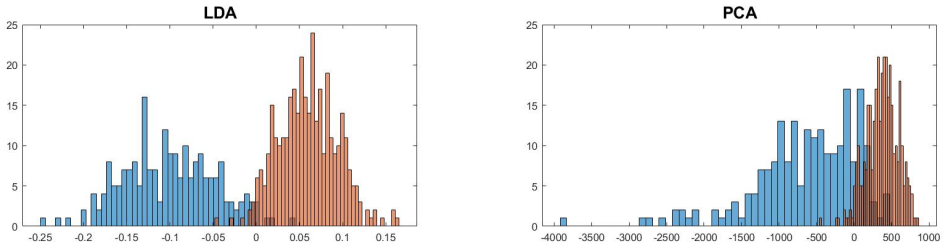

Figure 3: Problem 3: WisconsinBreastCancerData

3.3 (c)

I choose the larger components 6, 15, 17, 18, 20, 28 and 30 and remove them. The dimension is reduced from 30 to 7. As shown in Figure 4, the LDA result with reduced data still captures the differences between the clusters, but it's not better than the result with full data. It still can't seperate the two clusters completely, although PCA result is better. Although some attributes play less important roles in classifying, we can't overlook their effects.

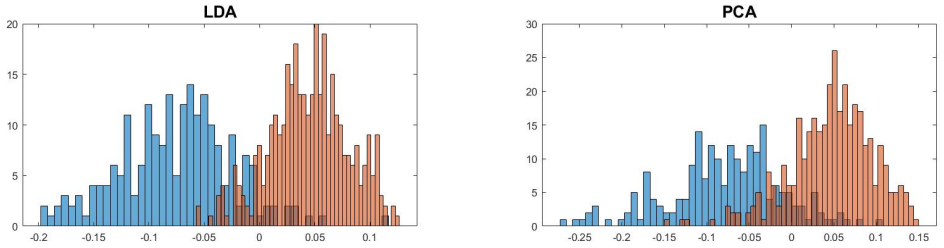

Figure 4: Problem 3: WisconsinBreastCancerData (reduced)

4 Problem 4

The result is shown in Figure 5. Based on the collected attributes, the three clusters are neatly separated.

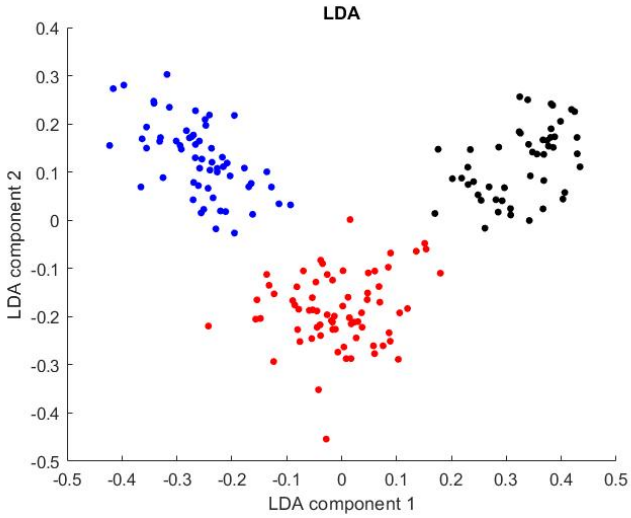

Figure 5: Problem 4: WineData

I choose the six largest components 1, 6, 7, 8, 11 and 12 and remove them. The dimension is reduced from 13 to 6. As shown in Figure 6, the three clusters are still seperated but the result is worse than above.

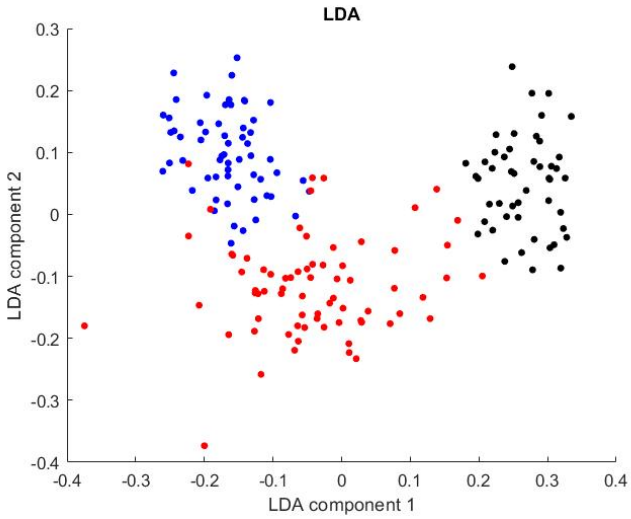

Figure 6: Problem 4: WineData (reduced)

5 Problem 5

Compute the three LDA separating vectors and plot them in Figure 7 and 8. Figure 7 shows LDA component 1,2. Figure 8 shows LDA component 1,3 and 2,3.

In Figure 7, the first two LDA directions separate the four groups well. While in Figure 8, both of the two pairs of LDA directions can't separate the groups clearly.

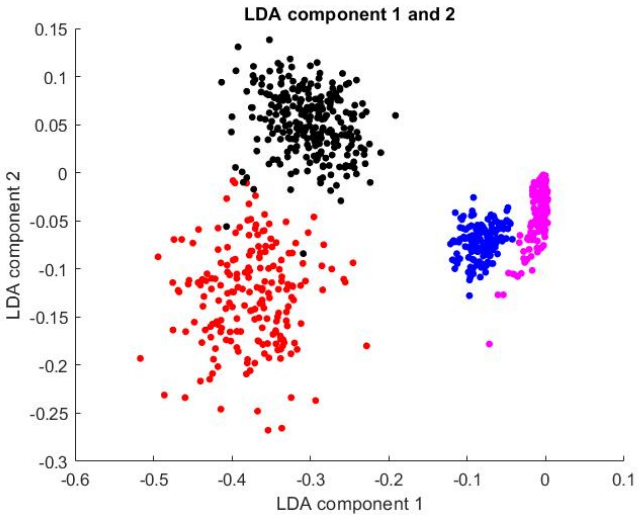

Figure 7: Problem 5: ForestSpectra

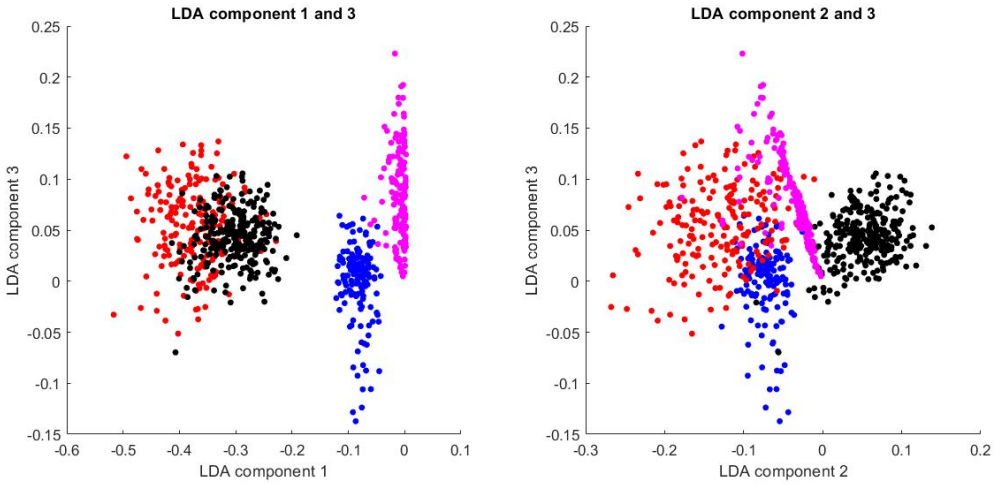

Figure 8: Problem 5: ForestSpectra

6 Matlab Code

6.1 LDA function

```
function [z,Q] = lda(k, X, I)
% Jiasen Zhang LDA

% input
% k = number of clusters
% X = data
% I = partition

% output
% z = LDA reduced data
% Q = seperating direction

[n,p]=size(X);

c=mean(X,2); % global mean
ck = zeros(n,k); % cluster means
S_w = zeros(n,n);
for L=1:k
    % cluster means
    ck(:,L) = mean(X(:,I==L),2);
    % compute S_L and S_w
    X_LC = X(:,I==L) - ck(:,L);
    % S_L = X_LC*X_LC';
    S_w = S_w + X_LC*X_LC';
end

% compute S_b
X_bar = zeros(n,p);
for j=1:p
    L = I(j); % cluster
    X_bar(:,j) = ck(:,L);
end
S_b = (X_bar-c)*(X_bar-c)';

% get matrix A
d1 = max(eig(S_w));
tau = 0; % if not positive definite, change to 1e-16
S_we = S_w + tau*d1*d1*eye(n);
K = chol(S_we);
A = inv(K)'*S_b/K;

% k-1 largest eigenvectors of A and solve Q
Q = zeros(n,k-1);
[v,d]=eig(A);
d = diag(d)/sum(diag(d)) % proportion of trace
fprintf('Proportion of trace:\n');
for j=1:k-1
    maxindex = find(d==max(d)); % find the maximum eigenvalue
    Q(:,j) = K\v(:,maxindex);
    fprintf('%.8f\n',d(maxindex)); % show proportions of first k-1 eigenvalues
    d(maxindex)=0;
end

% LDA reduced data
z = Q'*X;
end
```

6.2 Problem 1

```
clear all;clc;

% three clusters
I = ones(600,1);
I(201:400)=2;
I(401:600)=3;

% generate data X
X = zeros(3,600);
X(1,1:200)=normrnd(-3,1,1,200);
X(2,1:200)=normrnd(-3,1,1,200);
X(3,1:200)=normrnd(-3,1,1,200);
X(1,201:400)=normrnd(2,1,1,200);
X(2,201:400)=normrnd(0,1,1,200);
X(3,201:400)=normrnd(0,1,1,200);
X(1,401:600)=normrnd(4,1,1,200);
X(2,401:600)=normrnd(4,1,1,200);
X(3,401:600)=normrnd(-6,1,1,200);
cov(X(:,1:200)')
cov(X(:,201:400)')
cov(X(:,401:600)')
Xbar = mean(X,2);
Xc = X - Xbar;

k = 3;
[z,Q] = lda(k, X, I);

% plot
scatter(z(1,I==1),z(2,I==1),'b','k.','SizeData',200);hold on;
scatter(z(1,I==2),z(2,I==2),'r','k.','SizeData',200);
scatter(z(1,I==3),z(2,I==3),'k','k.','SizeData',200);
xlabel('LDA component 1');
ylabel('LDA component 2');
title('LDA');
```

6.3 Problem 2

```
clear all;clc;

% Given data and partition
load IrisData
[n,p]=size(X);
I = zeros(1,p);
I(1:50)=1;
I(51:100)=2;
I(101:end)=3;
Xbar = mean(X,2);
Xc = X - Xbar;

k = 3;
[z,Q] = lda(k, Xc, I);

subplot(1,2,1);
scatter(z(1,I==1),z(2,I==1),'b','k.','SizeData',200);hold on;
scatter(z(1,I==2),z(2,I==2),'r','k.','SizeData',200);
scatter(z(1,I==3),z(2,I==3),'k','k.','SizeData',200);
xlabel('LDA component 1');
ylabel('LDA component 2');
title('LDA');

[u,¬,¬] = svd(Xc);
z = u(:,1:2)'*Xc;
subplot(1,2,2);
scatter(z(1,I==1),z(2,I==1),'b','k.','SizeData',200);hold on;
```

```
scatter(z(1,I==2),z(2,I==2),'r','k.','SizeData',200);
scatter(z(1,I==3),z(2,I==3),'k','k.','SizeData',200);
xlabel('Principal component 1');
ylabel('Principal component 2');
title('PCA');
```

6.4 Problem 3

```
clear all;clc;

% Given data and partition
load WisconsinBreastCancerData
X = Data_WCD_Matrix;
I = I_Label;
[n,p]=size(X);
Xbar = mean(X,2);
Xc = X - Xbar;

k = 2;
[z,Q] = lda(k, Xc, I);

subplot(1,2,1);
histogram(z(I==1),50);hold on;
histogram(z(I==2),50);
title('LDA','FontSize',16);
subplot(1,2,2);
[u,¬,¬] = svd(Xc);
z0 = u(:,1)'*Xc;
histogram(z0(I==1),50);hold on;
histogram(z0(I==2),50);
title('PCA','FontSize',16);

% (c)
% select the largest components of Q:
% component 6,15,17,18,20,28,30
choose = [6,15,17,18,20,28,30];
Xc2 = Xc(choose,:);

% repeat
[z,Q0] = lda(k, Xc2, I);
figure();
subplot(1,2,1);
histogram(z(I==1),50);hold on;
histogram(z(I==2),50);
title('LDA','FontSize',16);
subplot(1,2,2);
[u,¬,¬] = svd(Xc2);
z0 = u(:,1)'*Xc2;
histogram(z0(I==1),50);hold on;
histogram(z0(I==2),50);
title('PCA','FontSize',16);
```

6.5 Problem 4

```
clear all;clc;

load WineData
Xbar = mean(X,2);
Xc = X - Xbar;

k = 3;
```

```
[z,Q] = lda(k, Xc, I);

scatter(z(1,I==1),-z(2,I==1),'b','k.','SizeData',200);hold on;
scatter(z(1,I==2),-z(2,I==2),'r','k.','SizeData',200);
scatter(z(1,I==3),-z(2,I==3),'k','k.','SizeData',200);
xlabel('LDA component 1');
ylabel('LDA component 2');
title('LDA');

% select the largest components:
% component 1,6,7,8,11,12
choose = [1,6,7,8,11,12];
Xc2 = Xc(choose,:);
[z,Q2] = lda(k, Xc2, I);

figure();
scatter(z(1,I==1),z(2,I==1),'b','k.','SizeData',200);hold on;
scatter(z(1,I==2),z(2,I==2),'r','k.','SizeData',200);
scatter(z(1,I==3),z(2,I==3),'k','k.','SizeData',200);
xlabel('LDA component 1');
ylabel('LDA component 2');
title('LDA');
```

6.6 Problem 5

```
clear all;clc;

load ForestSpectra
k = 4;
I=Itype;
[z,Q] = lda(k, X, I);

% LDA components 1,2
scatter(z(1,I==1),z(2,I==1),'b','k.','SizeData',200);hold on;
scatter(z(1,I==2),z(2,I==2),'r','k.','SizeData',200);
scatter(z(1,I==3),z(2,I==3),'k','k.','SizeData',200);
scatter(z(1,I==4),z(2,I==4),'m','k.','SizeData',200);
xlabel('LDA component 1');
ylabel('LDA component 2');
title('LDA component 1 and 2');
% LDA components 1,3 and 2,3
figure();
subplot(1,2,1);
scatter(z(1,I==1),z(3,I==1),'b','k.','SizeData',200);hold on;
scatter(z(1,I==2),z(3,I==2),'r','k.','SizeData',200);
scatter(z(1,I==3),z(3,I==3),'k','k.','SizeData',200);
scatter(z(1,I==4),z(3,I==4),'m','k.','SizeData',200);
xlabel('LDA component 1');
ylabel('LDA component 3');
title('LDA component 1 and 3');
subplot(1,2,2);
scatter(z(2,I==1),z(3,I==1),'b','k.','SizeData',200);hold on;
scatter(z(2,I==2),z(3,I==2),'r','k.','SizeData',200);
scatter(z(2,I==3),z(3,I==3),'k','k.','SizeData',200);
scatter(z(2,I==4),z(3,I==4),'m','k.','SizeData',200);
xlabel('LDA component 2');
ylabel('LDA component 3');
title('LDA component 2 and 3');
```