HW5
CS 273
Debatri Mitra

March 26, 2015

1 Problem -1

1.1

The first two featurs of iris data are loaded and scatter plot is shown here.

```
data = load('data/iris.txt');
X=data(:,1:2);
scatter(X(:,1),X(:,2));
```

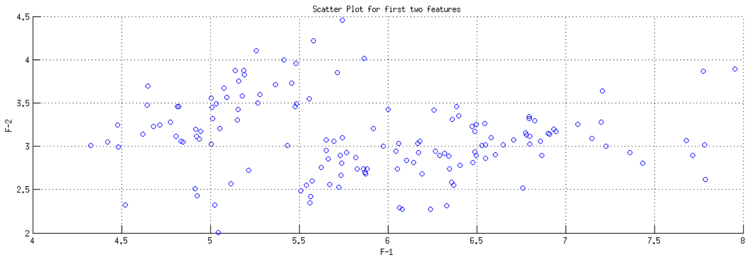

Figure 1: Data

1.2

K means trained for K=5, K=20. It has been run 25 times and the best score and plot reported.

```
[assign, clusters, sumd] = Kmeans(X,5,'k++'); % 5 clusters
figure; plotClassify2D([],X,assign); sumd
[assign, clusters, sumd] = Kmeans(X,20,'k++'); % 5 clusters
figure; plotClassify2D([],X,assign); sumd
```

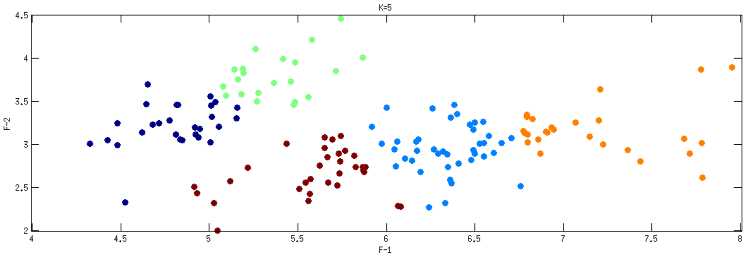

Figure 2: K=5

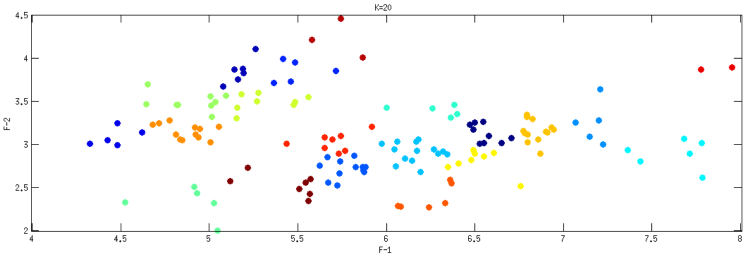

Figure 3: K=20

The best reported results are for K=5
21.3414
for K=20
4.1583
I think K=20 is too many clusters and also K=5 is not optimal based on the two features. Though as we know there are 5 species K=5 should be optimal but, using only two features K=2/3 seems more reasonable just looking at the data.

1.3

Aggomorative Clustering with K=5 and K=20 with both Single and Complete Linkage have been performed. Here is the code:

```
cluster = agglomCluster(X,5,'min'); % single linkage
figure; plotClassify2D([],X,cluster);
cluster = agglomCluster(X,5,'max'); % complete linkage
figure; plotClassify2D([],X,cluster);
cluster = agglomCluster(X,20,'min'); % single linkage
figure; plotClassify2D([],X,cluster);
cluster = agglomCluster(X,20,'max'); % complete linkage
figure; plotClassify2D([],X,cluster);
```

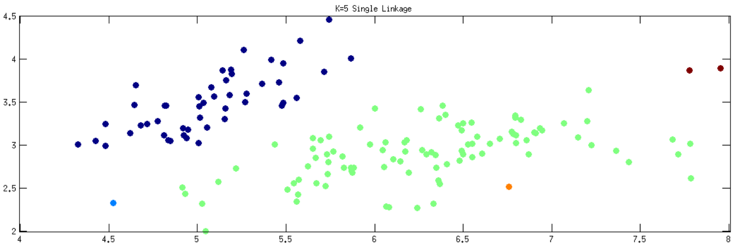

Figure 4: K=5 Single Linkage

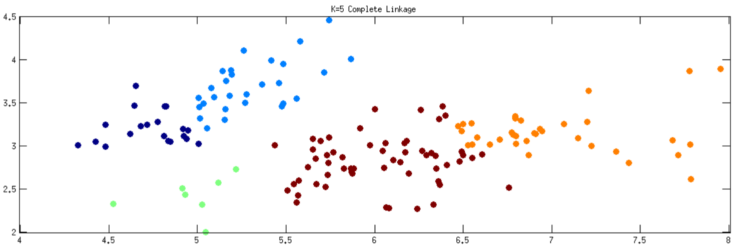

Figure 5: K=5 Complete Linkage

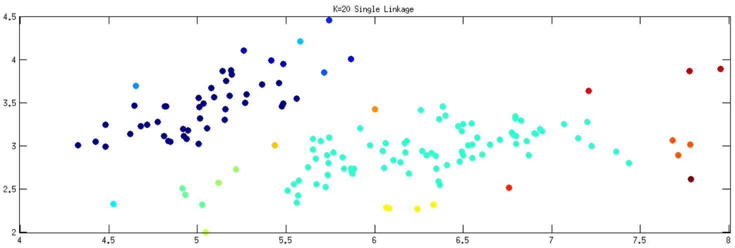

Figure 6: K=20 Single Linkage

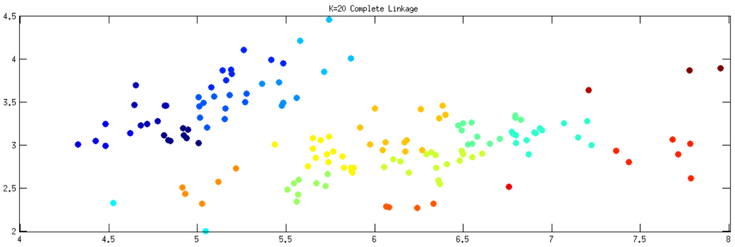

Figure 7: K=20 Complete Linkage

Single Linkage looks to give poor result whereas Complete linkage is better. K means performed similar to the results given by complete linkage.

1.4

EM for GMM has been performed using both K=5 and K=20. The results are displayed along with the Log Likelihood plots. It has been run 5 times and the best score has been reported.
for K=5 Log Likelihood= -205.4390 and for K=20 Log Likelihood is = -113.2676

```
[z,T,soft,ll]= emCluster(X,5,’k++’);
display(ll);
[z,T,soft,ll] = emCluster(X,20,’k++’); ll % 20 clusters
display(ll);
```

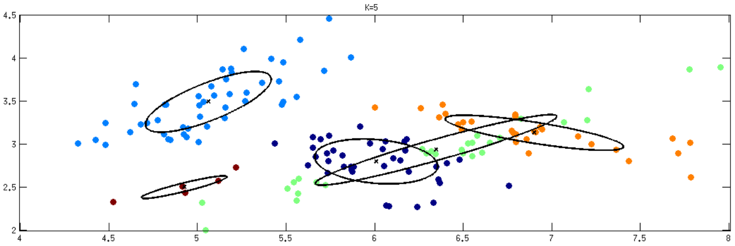

Figure 8: K=5 EM

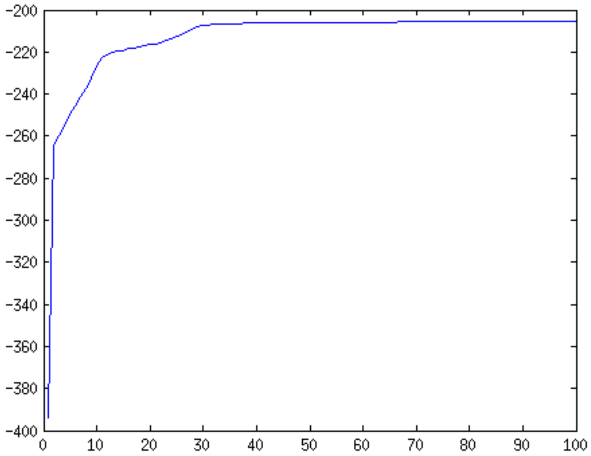

Figure 9: K=5 EM Log Likelihood

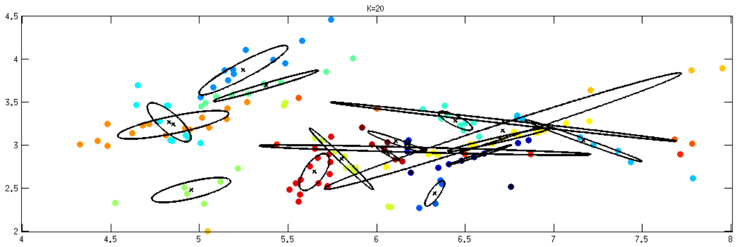

Figure 10: K=20 EM

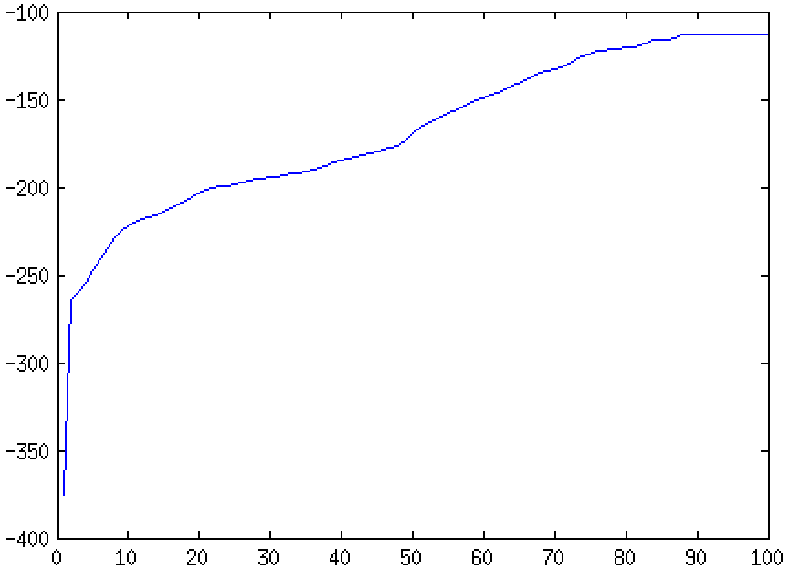

Figure 11: K=20 EM Log Likelihood

This does a much better job than either k means or agglomorative and finds hidden variables. However, for this simple data it looks like overkilling and not appropriate. For higher dimensional data this might be more useful. Here for K=5 though it looks OK Again K =20 is visibly too many clusters.

2 Problem-2

2.1

K - means performed using the provided function. The SSD found was 2. here is the code -

```
clear all
[vocab] = textread('vocab.txt','%s');
[did,wid,cnt] = textread('docword.txt','%d%d%d','headerlines',3);
X=sparse(did,wid,cnt);
D=max(did);
W=max(wid);
N=sum(cnt);
Xn= X./ repmat(sum(X,2),[1,W]) ;
[z,c,sumd] = Kmeans(Xn,20,'k++');
display(sumd);
```

2.2

The same code run for 5 instances and the best has been kept. The SSD obtained are - 2.0451,2.3477,2.1229,2.1229,2.1229,2.1229. So, the results from the Best 2.0451 is retained. Here is the code

```
z=zeros(202,1);
c=zeros(20,1914);
sumd=inf;

for i=1:5
    [z_i,c_i,sumd_i] = Kmeans(Xn,20,'k++');
    if (sumd_i<sumd)
        z=z_i;
        c=z_i;
        sumd=sumd_i;
    end
    display(sumd);
end
```

2.3

The Cluster frequencies are found. Here is the cluster frequencies
2 ,1, 1, 1, 1, 2, 40, 4, 3, 3, 1, 14, 2, 110, 1, 1, 1, 3, 2, 9

```
H=hist(z,1:20);
hist(z,1:20)
```

Here is the plot

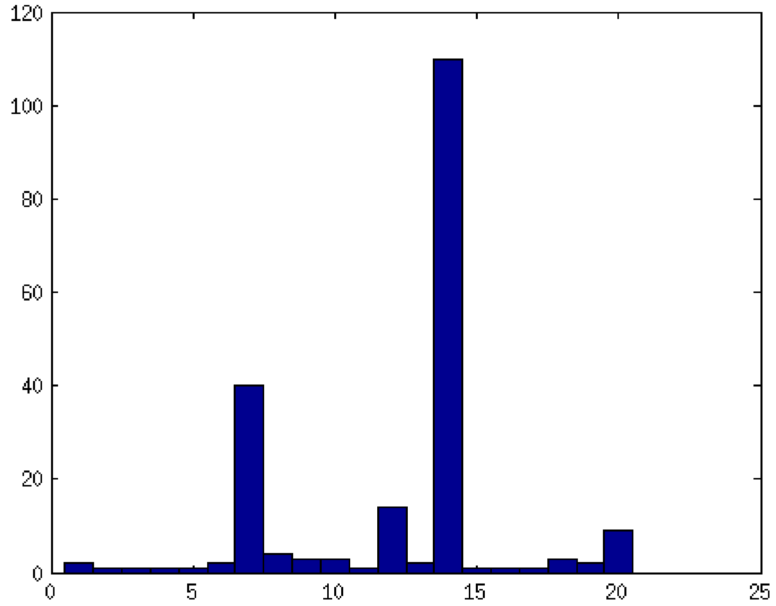

Figure 12: Cluster Frequencies

2.4

TOP 10 words from each cluster center

1: feet square broadway seventh street side times avenue block million
2: bradley campaign candidates mccain national hampshire bush marijuana drug political
3: season rusie giants game games winner 1900 career century entire
4: fff xff fxf ffx xfx ffff fffx ffxf tomjanovich fxx
5: vick quarterback tech florida game coach season syracuse virginia yards
6: white mall president guests house clinton dinner crowd millennium room
7: economy government putin system america businesses country economic president russia
8: bishop archbishop cardinal buffalo chicago church close late pittsburgh seattle
9: y2k 2000 computer problem problems saturday koskinen systems computers system
10: cannibalism human eating ago 000 turner university evidence french remains
11: american war europe algeria manas political country epic boot language
12: millennium times square 2000 city midnight night york 000 friday
13: putin yeltsin russian russia grozny president power chechnya troops kremlin
14: century book amp week finds lives war school boy scholastic

15: fireworks calls celebrations city island midnight night times air began
16: team games nba percent coach going line lot orlando things
17: team game season players test games coach end home league
18: bowl yards game stanford quarter point rose wisconsin minutes set
19: yeltsin russia chechnya government putin reform russian democracy economic political
20: central snow states valley weather week air band century early

some form interpretable group. 1 = roads 2 = political 3 = baseball 5= game 7= politics ....and so on.

2.5

z(1), z(15), z(30) The associated clusters are 5,12,14

for z(15):

Despite a few sputters and glitches, the world's computers appear to have survived the Year 2000 rollover without major

The eastern half of the United States smoothly followed the rest of the world into the new century early Saturday

Now that the Year 2000 has arrived around the world without significant disruptions of power, transportation, commerce and

It was three hours before midnight on Friday and John A. Koskinen did not appear worried a bit. His head was back.

Before the countdown, the cork popping, the confetti and the computer worries indeed, long before most people had

As the world glided smoothly into the new century, the United States reported Saturday only a smattering of minor Y2K

A stockpile of 750,000 worth of spare parts set aside for Year 2000 fixes remained in a storeroom Saturday in New

Before the countdown, the cork popping, the confetti and the computer worries indeed, long before most people had

Moving in (w) Washington and (f) financial categories. By MARILYN GEEWAX

Before the countdown, the cork popping, the confetti and the computer worries indeed, long before most people had

As the world glided smoothly into the new century, the United States reported Saturday only a smattering of minor Y2K

So, was Y2K the most overhyped phenomenon in 2,000 years? Or did the diligence of a world preparing as if for war stave

This looks lik Y2K, computers z(30)=

For those who believe that in the good old days before calculators, before computers people were better at mental

EDS: New top and updates throughout. By SCOTT MONTGOMERY

Airports around the world, ordinarily quiet on New Year's Eve, were unusually so Friday as thousands of potential travelers stayed

Ray Hubbard, a television producer and broadcasting executive and a pioneer in the medium, died on Dec. 27 in Kenwood, Calif. He

THIRD MILLENNIUM CELEBRATIONS CAST LIGHT ON TWO MEXICOS By FRANC CONTRERAS

Two thousand years after Christ's obscure birth in a dusty town in Judea, the world's 6 billion people most of them non-Christian

The millennium, an idea with overtones ranging from Biblical to commercial, had swelled recently into a coercive miniculture as the

The festivities for most of India's 1 billion people were muted by comparison with those in wealthier nations, but hundreds of

Rain and chilly weather didn't keep thousands of paradegoers from camping out Friday night for the 111th Tournament

Although it had seen many huge crowds over the years, for inaugurations, Fourth of July fireworks, protests and

How do you mark the passage of 1,000 years of human history? Simple, in this city of chronic gamblers and cold-eyed

Don't call Jim and Susan Smith survivalists. In fact, you don't even have to call them Jim and Susan those

Looks like 4th of July crowd independence day

z(1)=

School has been out at Cal State Northridge since the week before Christmas, but since you can learn something every

Fred Saigh, a former owner of the St. Louis Cardinals who was forced out of baseball in 1953 when he was sentenced to federal

Back in the dog days of August, nine victories, a championship in the southern half of the Big 12, and an encore New

Texas' off-season of discontent began when the final gun sounded at Saturday's Southwestern Bell Cotton Bowl, sealing a

The total of six yards rushing in the Big 12 championship loss to Nebraska was a school worst. But the minus-27 yards rushing in

It would have been easy. Very easy. Rudy Tomjanovich had his championships. He was the next Dream

The 86th Rose Bowl was billed as a potential offensive shootout, but it turned into a defensive battle decided

Who could know? Oh, they had a hunch when he came out of high school, with lots

At halftime of Boston College's debacle.com against the University of Colorado Friday in the Insight.com Bowl, coach Tom

Motivation was never an issue. To keep focused on rehabilitating a stress fracture in his right foot, Danny Fortson

Eddie Jones has rejoined the Charlotte Hornets and reaffirmed his plans for a quick comeback from an elbow injury.

Just four months ago, only those living in the small farming town tucked between the Blue Ridge Mountains and

looks like school Sports related

2.6

Repeating verything for K=40 here are the SSD
1.7285, 1.6765, 1.7701, 1.6992, 1.6866 The clustr assignmnts look lik this: The 40 clusters look like this: 1: children including post produced programs television american art black calif

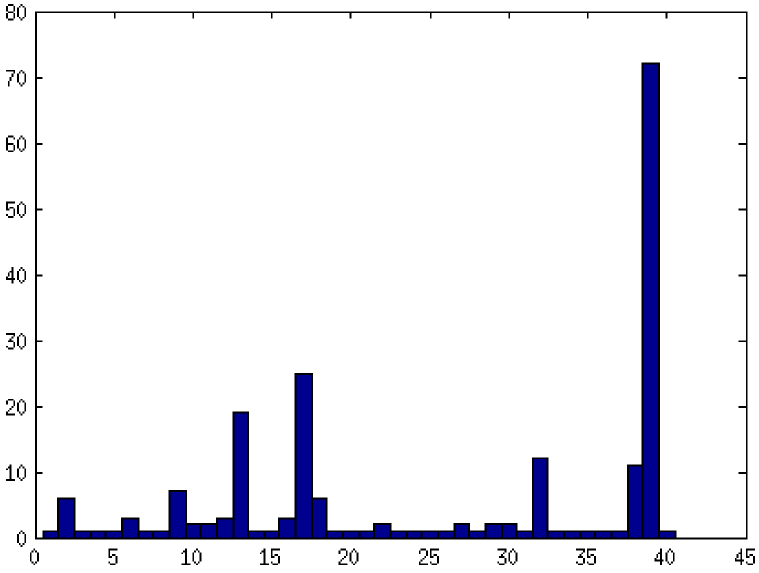

Figure 13: Cluster Frequencies

2: 2000 problem computer y2k systems system friday officials computers power
3: jackson game lakers star conference practice coach coaches fun games
4: marks cooking food writing book africa brooklyn business foreign getting
5: pakistan hijackers border hostages india government headed hijacking ended hours
6: hijackers hostages government indian burger plane india passengers told deal
7: concert tickets 500 center didn eagles missed party waited wife
8: combs million harrison john matter police pulled say star attention
9: buses greene authority diesel city natural gas company plan york
10: fireworks calls celebrations city island midnight night times air began
11: algeria islamic war americans country army europe front independence conflict
12: bowden florida coach football virginia seminoles tech bowl team big
13: y2k 2000 government news millennium russia saturday putin president york
14: test end houston 000 0101 0102 100 1900 1900s 1968
15: mississippi play scored game half points ranked thomas night team

16: y2k problems 2000 saturday koskinen computer going computers bug spent
17: square times millennium city 2000 000 midnight crowd night party
18: yeltsin putin russia russian political chechnya president power prime russians
19: night parade rain family millennium eve friends want calif colorado
20: games million federal spent city olympic olympics atlanta cities commission
21: island celebration filled fireworks lot millennium added americans beach beat
22: marijuana drug drugs nadelmann policy criminal director foundation group americans
23: issue number days times 500 front paper 000 april desk
24: season rusie giants game games winner 1900 career century entire
25: news cnn abc hours millennium clinton midnight coverage entertainment friday
26: warrick heisman florida award guy national championship college field game
27: communications health amp american book called century recalls western wrote
28: cats beijing owners police association called carry chinese eat eating
29: tutsi hutu rwanda burundi ethnic country experts africa van 1994
30: buildings architecture architects landmark modern church national columbus historic foundation
31: casey coach fassel nets van home jersey parcells garden knicks
32: bradley candidates mccain campaign hampshire bush political republican president voters
33: internet sales commerce services customers holiday information small web agents
34: opened bank according account cash country early money 2000 banks
35: plummer dictionary savannah school lady book community daughter 000 began
36: fortson game celtics play denver boston games foot forward going
37: completed dancers cast director ends stage thomas 1995 advantage afternoon
38: season team league players coach game giants games teams football
39: game team century season games players york going sports million
40: additional accounts trust computer account addition date related business department

Here the assignments look similar for 1,15,30 ...however, breaks down and more refined.

3 Problem 3

3.1

```
X = load('data/faces.txt');
```

```
% load face dataset
%img = reshape(X(i,:),[24 24]); % convert vectorized datum to 24x24 image patch
%imagesc(img); axis square; colormap gray;
mu = mean(X);
X0 = X - repmat(mu,[size(X,1),1]);
```

3.2

Taking the SVD

```
 [U S V] = svds(X0,25);
W = U*S;
```

3.3

For K=1,...,20 the Reconstruction errors have been calculated and plotted as a function of K

```
  for k=1:10,
 X_predict =W(:,1:k)*V(:,1:k)';
E(k) = mean(mean( (X0 - X_predict).^2 ));
end;
 figure; plot(1:10, E);
```

3.4

First 3 principal directions of the data computing + and - from μ by scaling factor α as given in the problem. The result is shown here

```
 for k=1:3,
 alpha = 2*median(abs(W(:,k)));
A = reshape(mu + alpha*V(:,k)', [24 24]);
B = reshape(mu - alpha*V(:,k)', [24 24]);
 figure;
 subplot(2,1,1);imagesc(A); colormap gray;
 subplot(2,1,2);imagesc(B); colormap gray;
end;
```

3.5

I chose first 30 . Here is the code

```
idx = [1:30];
% pick some data
 figure; hold on; axis ij; colormap(gray);
 range = max(W(idx,1:2)) - min(W(idx,1:2)); % find range of coordinates to be plo
 scale = [200 200]./range;
 for i=idx,
```

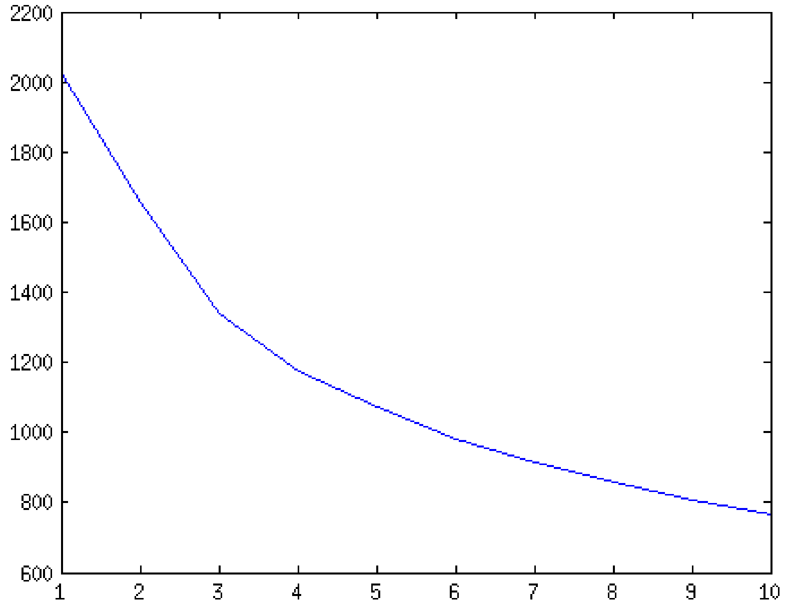

Figure 14: Error Plot

```
    imagesc(W(i,1)*scale(1),W(i,2)*scale(2), reshape(X(i,:),24,24));
end;
```

The Latent space representation is shown in the following fig.

3.6

Two images are reconstructed using K=1,10,25 The originals are here
CODE:

```
for i=[1 30],
im = X(i,:);
im = reshape(im, [24 24]); imagesc(im); colormap gray;
for k=[5 10 25],
im = mu+W(i,1:k)*V(:,1:k)';
im = reshape(im, [24 24]);
figure;
imagesc(im); colormap gray;

end;
end;
```

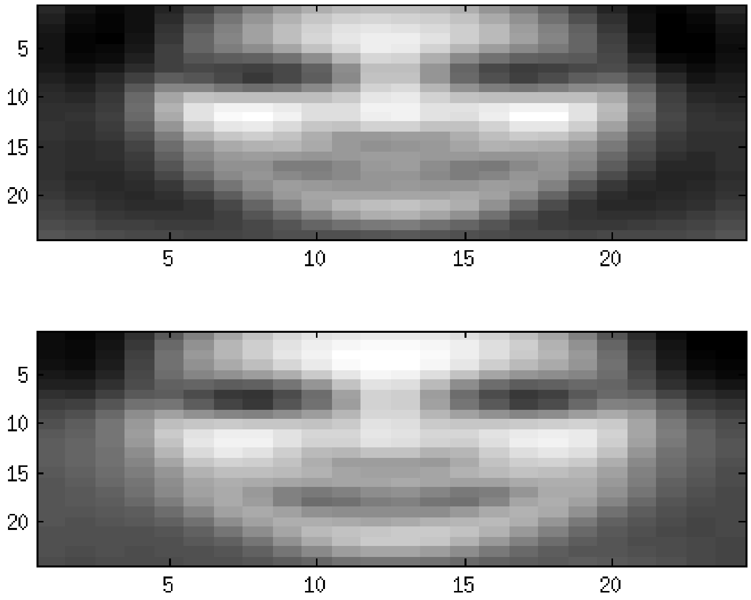

Figure 15: K=1

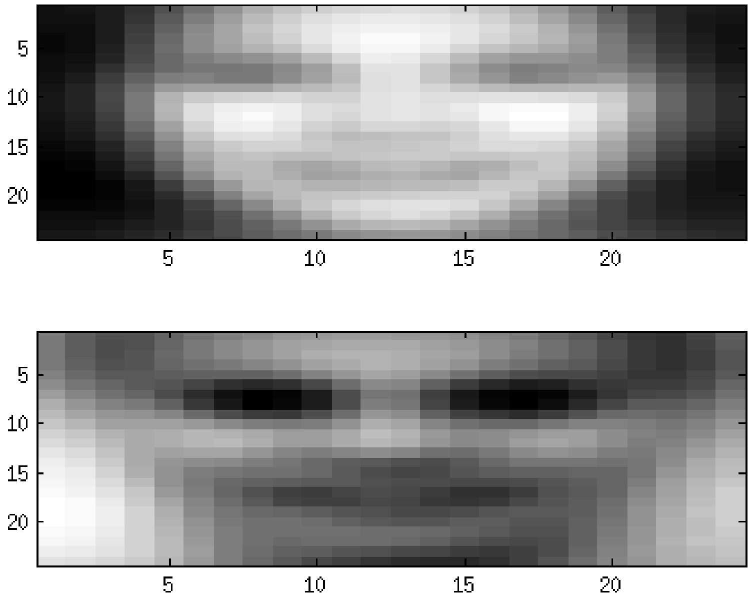

Figure 16: K=2

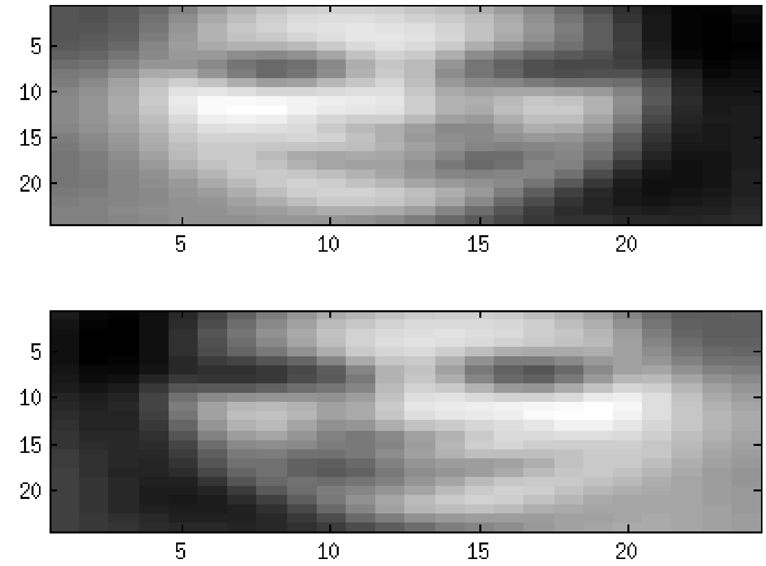

Figure 17: K=3

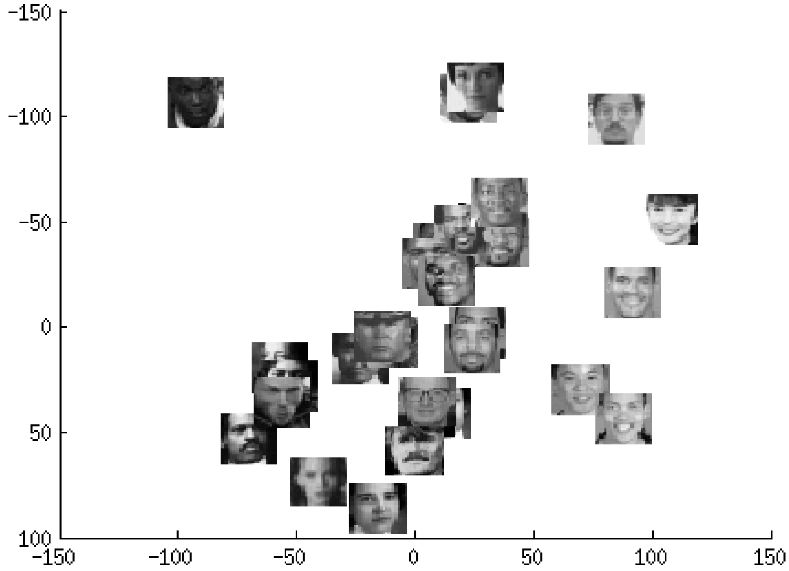

Figure 18: Latent Space Modelling

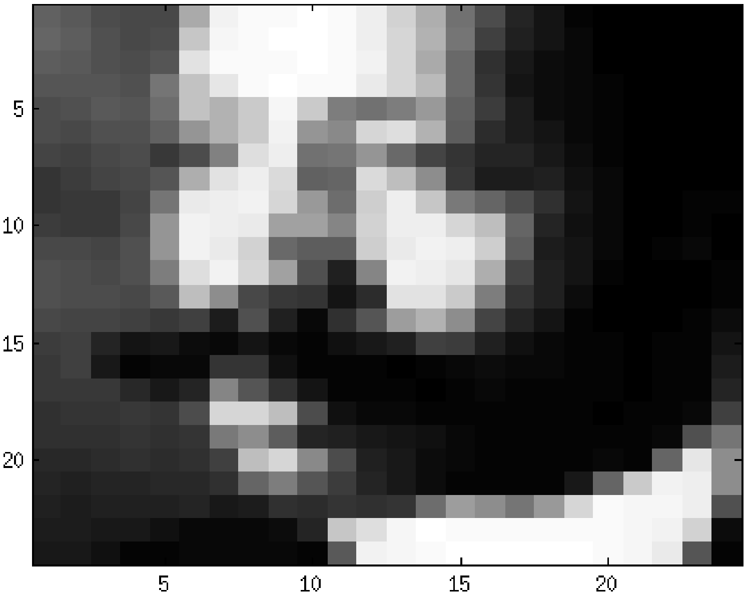

Figure 19: Original 1

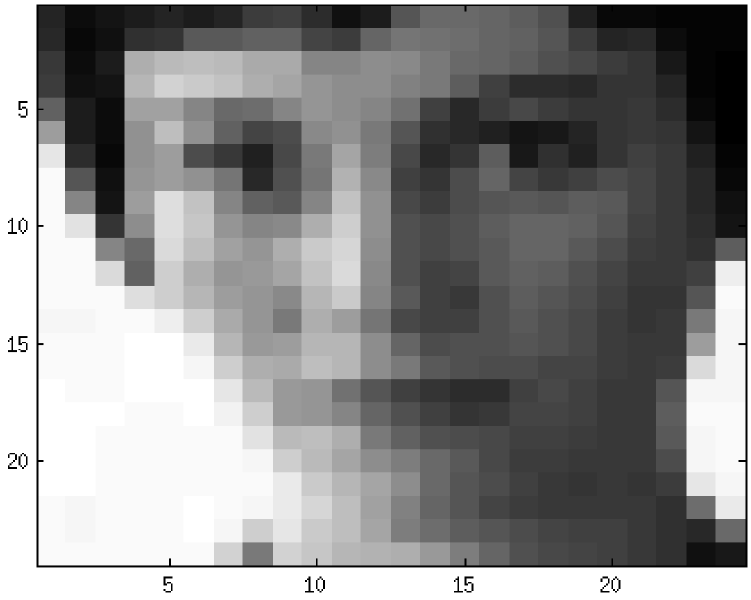

Figure 20: Original2

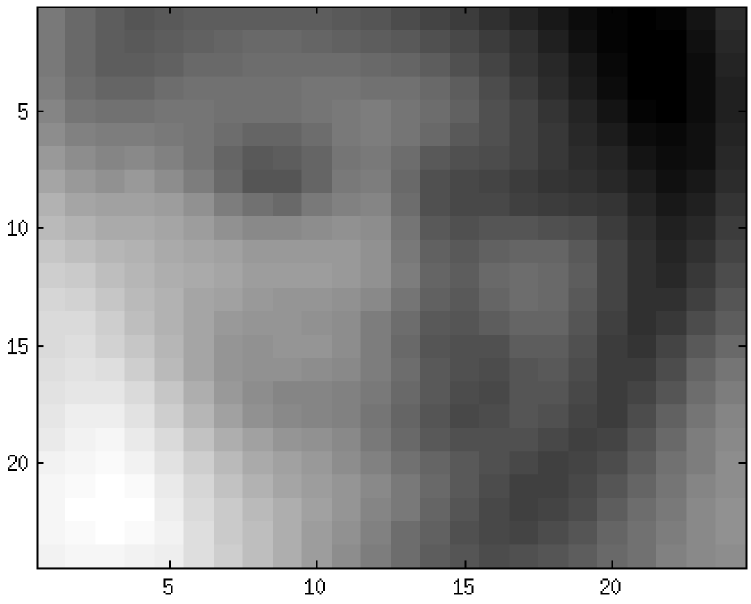

Figure 21: K=1 - Reconstruct Im2

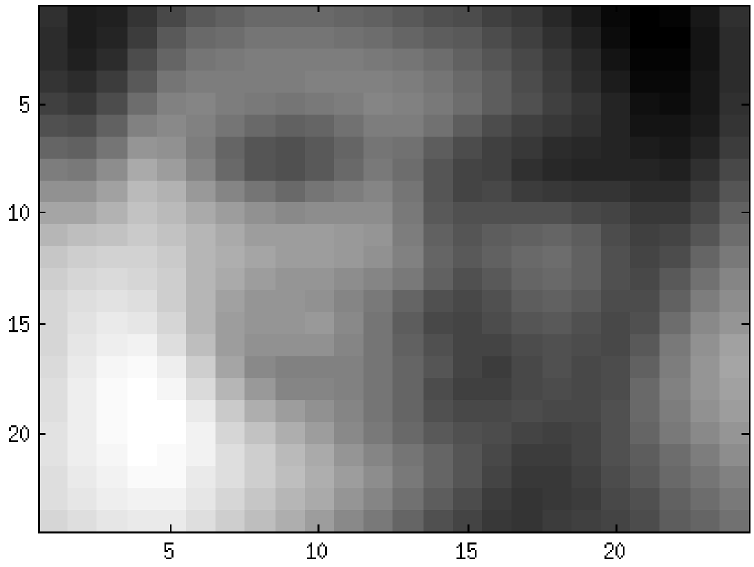

Figure 22: K=10 Reconstruct Im 2

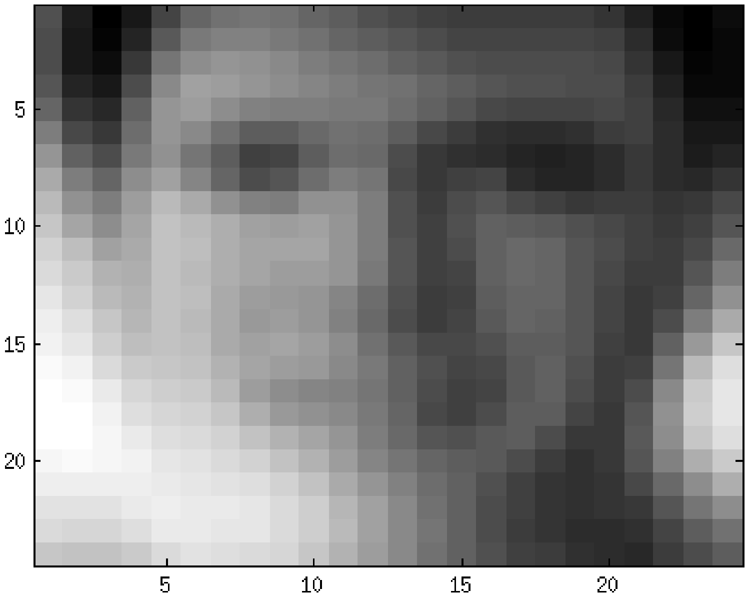

Figure 23: K=25 Reconstruct Im 2

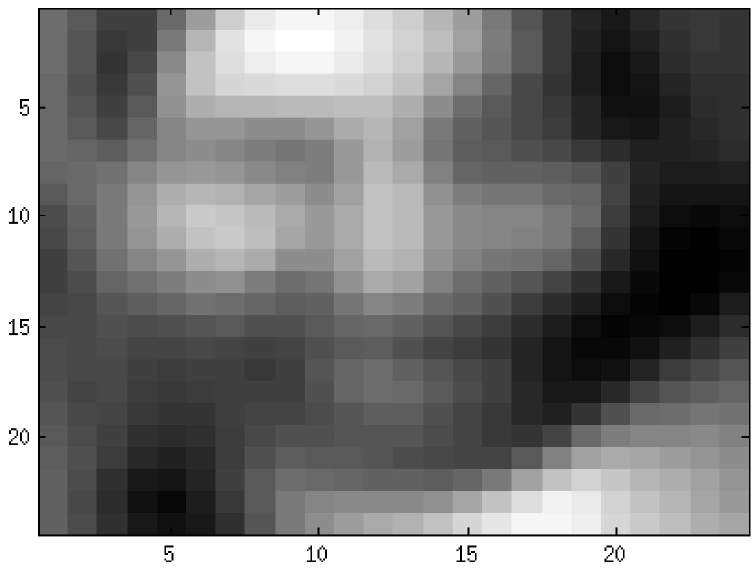

Figure 24: K=1 - Reconstruct Im1

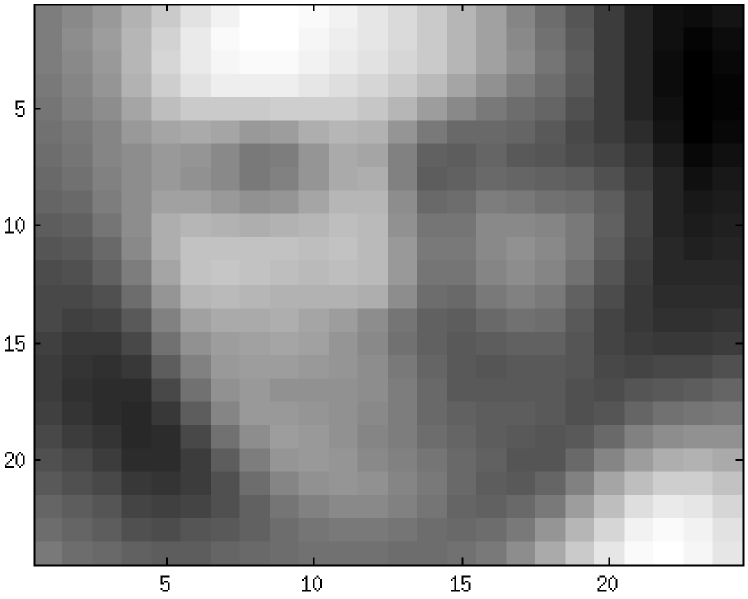

Figure 25: K=10 Reconstruct Im 1

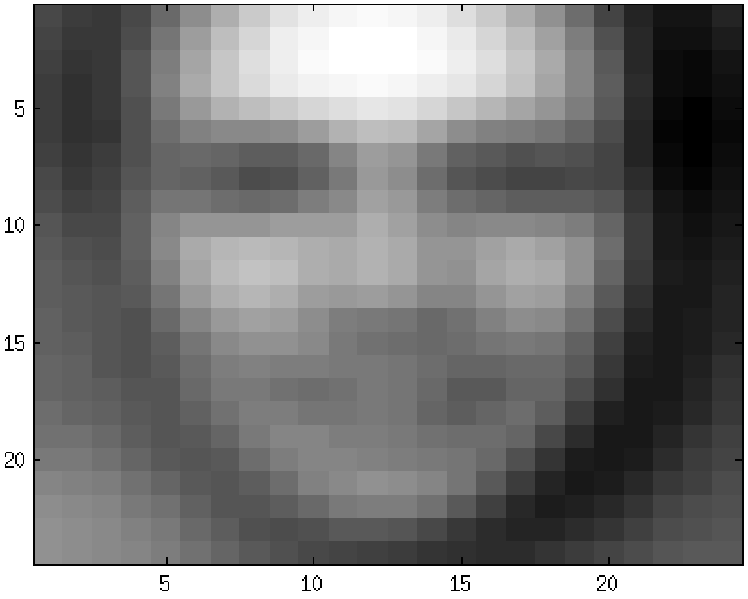

Figure 26: K=25 Reconstruct Im 1