Omer Solakli

2471677864

EE 511: Simulation of Stochastic Processes

Spring 2018

Project#3

1.

[Testing Faith]

```
num = xlsread('oldfaithful.xlsx');
Er=num(:,2);
Du=num(:,3);

scatter(Er,Du)
hold on

xlabel('eruption time mins');
ylabel('duration time mins');
X=[Er, Du];
opts = statset('Display','final');
[idx,C] = kmeans(X,2,'Distance','sqeuclidean',...
    'Replicates',5,'Options',opts);
figure;
plot(X(idx==1,1),X(idx==1,2),'r.','MarkerSize',12)
hold on
plot(X(idx==2,1),X(idx==2,2),'b.','MarkerSize',12)
plot(C(:,1),C(:,2),'kx',...
     'MarkerSize',15,'LineWidth',3)
legend('Cluster 1','Cluster 2','Centroids',...
       'Location','NW')
title 'Cluster Assignments and Centroids'
hold off
```

Above code is for the scatter plot of the 2-D data where ‘Er’ represents the Eruption time of the volcano and ‘Du’ represents the corresponding eruption duration. First, we read the data using ‘xlsread’ function of Matlab and generate scatter plot using the respective function. The scatter plot is shown below.

When we look at the scatter plot, we can easily see that data is already seems to be centered around to two clusters. One is with eruption time around 2 and duration around 55, and the other one is 4.5 and 80.

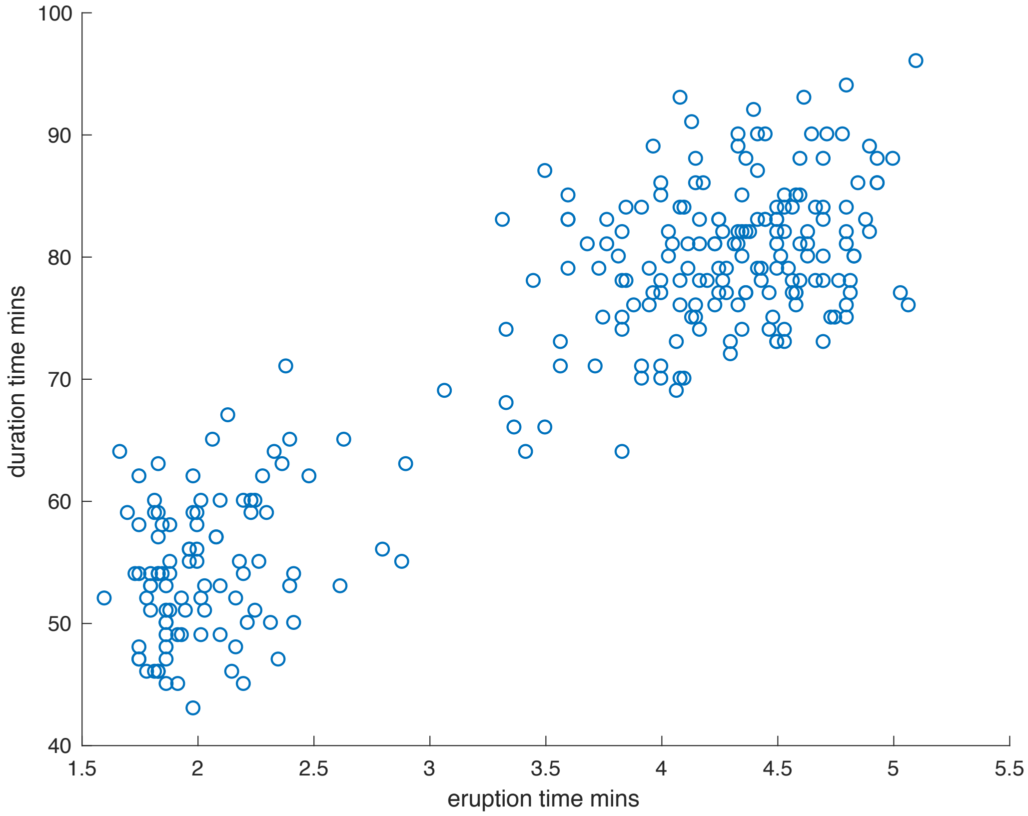

Figure 1: Scatter Plot

After we run k-means clustering algorithm with k=2. We get the following scatter plot. The distances from the cluster centres to the data points calculated using l2 norms, or Euclidian distances.

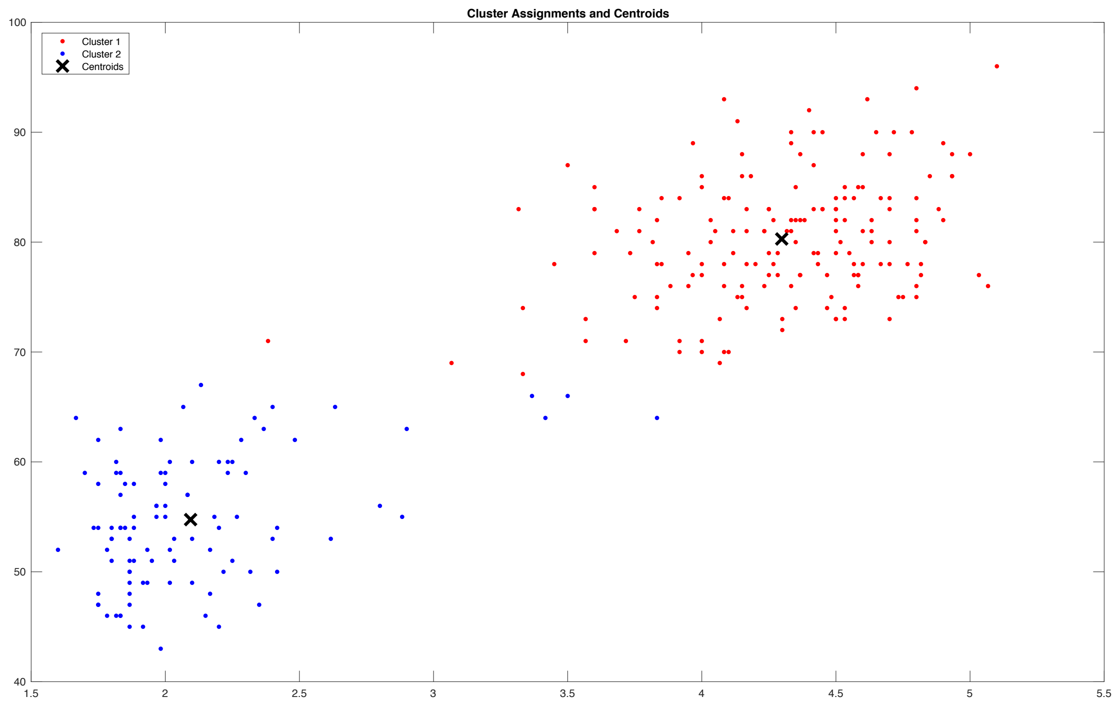

Figure 2: 2-means clustered scatter plot

The respective cluster centers are calculated as:

C =

4.2979 80.2849 [cluster center 1]
2.0943 54.7500 [cluster center 2]

2.
[EM]

```
%-------------------a--------------------%

mu = [0 5;4 -4]; %generating a mean matrix, where each row represents the
mean of two different gaussian distribution
sigma = cat(3,[2 .5],[1 1]); % 1-by-2-by-2 array
gm = gmdistribution(mu,sigma); %generates gaussian mixture distribution
with given mu and sigma, using 0.5 mixing probabilities.
ezsurf(@(x,y)pdf(gm,[x y]),[-10 10],[-10 10]) %surface plot

%-------------------a--------------------%
```

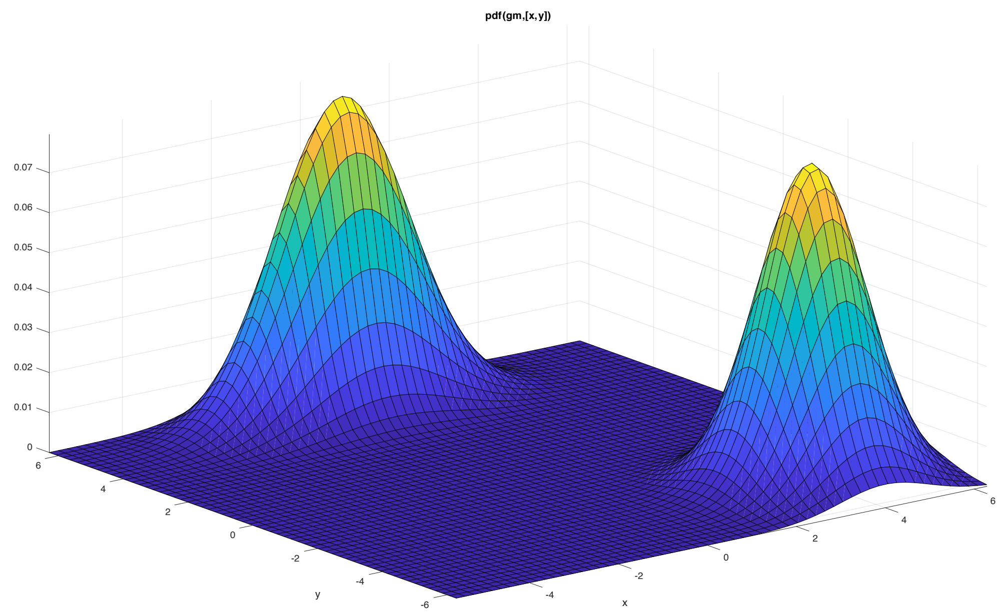

Figure 3: 2D GMM Distribution with p=0.5

Above 2D Gaussian Mixture Pdf is plotted.
Gaussian mixture distribution with 2 components in 2 dimensions
Component 1:
Mixing proportion: 0.500000
Mean: 0 5

Component 2:
Mixing proportion: 0.500000
Mean: 4 -4

One can randomize the mean and sigma matrices to generate random Gaussian mixture distribution.

```
%-------------------b-------------------%

mu1 = [1 2];            % Mean of the 1st component
sigma1 = [2 0; 0 .5]; % Covariance of the 1st component
mu2 = [-3 -5];          % Mean of the 2nd component
sigma2 = [1 0; 0 1];  % Covariance of the 2nd component


rng('default') % For reproducibility
r1 = mvnrnd(mu1,sigma1,1000); %generating random numbers using respective
mu and sigmas, chosen from multivaraite normal disribution
r2 = mvnrnd(mu2,sigma2,1000);
X = [r1; r2]; %putting the numbers in to the Matrix X
gm = gmdistribution.fit(X,2); %fits the distribution using Expectation
maximization algorithm to construct gm distribution object containing
maximum likelihood estimates of the parameters

scatter(X(:,1),X(:,2),10,'.') % Scatter plot with points of size 10
hold on
ezcontour(@(x,y)pdf(gm,[x y]),[-8 6],[-8 6]) %plotting the level curves on
the scatter plot

%-------------------b-------------------%
```

Above code uses EM algorithm to fit randomly generated numbers for GMM to fit the numbers to GMM distribution.

The mean for the first component is [1 2], whereas mean for the second component is [-3 -5]. After generating the numbers and fitting them to the distribution, we can check how algorithm estimates the parameters. The relevant output is shown below.

Gaussian mixture distribution with 2 components in 2 dimensions
Component 1:
Mixing proportion: 0.500000
Mean: -2.9617 -4.9727

Component 2:
Mixing proportion: 0.500000
Mean: 0.9539 2.0261

As one can see, algorithm estimates the parameters within (0-5)% error.

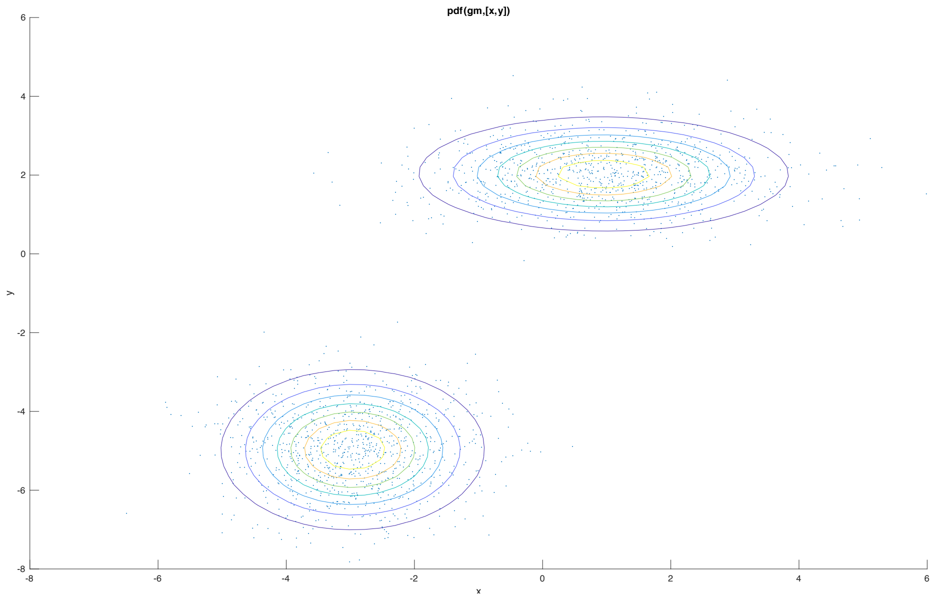

Figure 4: Scatter plot with Level surfaces for GMM

Above, is the scatter plot for the GMM model.

```
%------------------c-------------------%
mu1 = [1 2];           % Mean of the 1st component
sigma1 = [2 0; 0 2]; %  Spherical Covariance of the 1st component
mu2 = [-3 -5];         % Mean of the 2nd component
sigma2 = [1 0; 0 1];  % Spherical Covariance of the 2nd component

rng('default') % For reproducibility
r1 = mvnrnd(mu1,sigma1,300);
r2 = mvnrnd(mu2,sigma2,300);
X = [r1; r2];
tic
gm = gmdistribution.fit(X,2);
toc
scatter(X(:,1),X(:,2),10,'.') % Scatter plot with points of size 10
hold on
ezcontour(@(x,y)pdf(gm,[x y]),[-8 6],[-8 6])

%------------------c-------------------%
```

Elapsed time is 0.018139 seconds.

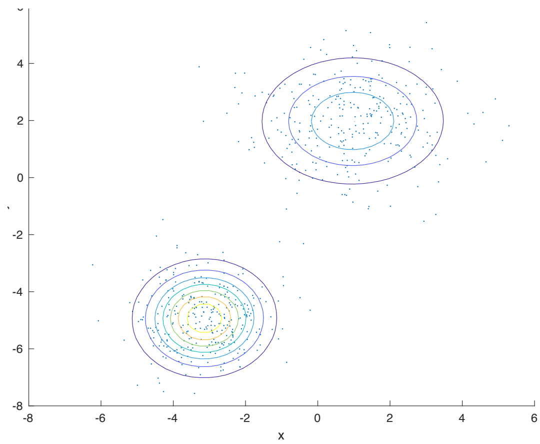

Figure 5: Plot for spherical covariances

```
mu1 = [1 2];             % Mean of the 1st component
sigma1 = [2 0; 0 10]; %  Ellipsoid Covariance of the 1st component
mu2 = [-3 -5];           % Mean of the 2nd component
sigma2 = [10 0; 0 1];  % Ellipsoid Covariance of the 2nd component
```

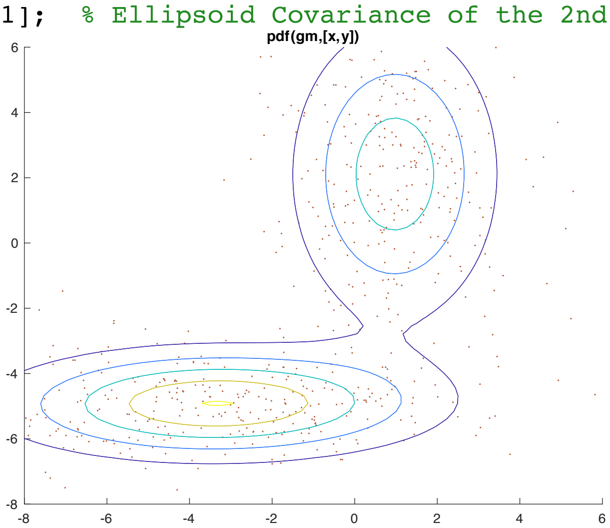

Figure 6: plot for ellipsoid covariances

Elapsed time is 0.003720 seconds.

```
mu1 = [0 -2];             % Mean of the 1st component (poorly seperated)
sigma1 = [2 0; 0 2]; %  Spherical Covariance of the 1st component
mu2 = [0 1];            % Mean of the 2nd component (poorly seperated)
sigma2 = [1 0; 0 1];  % Spherical Covariance of the 2nd component
```

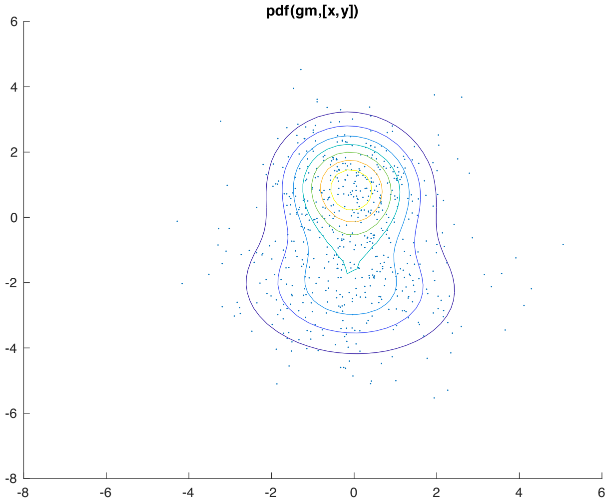

Figure 7: plot for poorly seperated data

Elapsed time is 0.048846 seconds.

We can see that the algorithm works fastest on ellipsoid covariance matrices, whereas it performs worst with poorly separated data.

```
%-----------------d-------------------%
num = xlsread('oldfaithful.xlsx');
Er=num(:,2);
Du=num(:,3);
X=[Er, Du];
gm = gmdistribution.fit(X,2);
figure
scatter(X(:,1),X(:,2),'.')
ezcontour(@(x,y)pdf(gm,[x y]),[0 10],[0 100])

%------------------d-------------------%
```

Gaussian mixture distribution with 2 components in 2 dimensions
Component 1:
Mixing proportion: 0.355885
Mean: 2.0364   54.4788

Component 2:
Mixing proportion: 0.644115
Mean: 4.2897   79.9684

As we can see, when we fit the data we can still estimate means and covariance's of 2 different clusters within a low error margin.

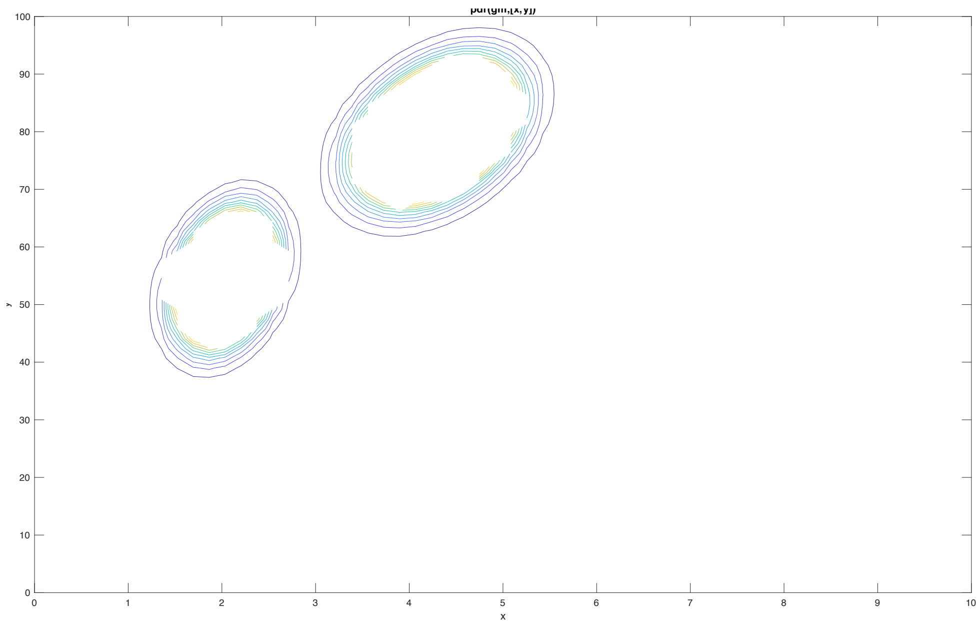

Figure 8: old faithful fit into GMM

3.
[Clusters of Text]

Below vector is the sum of distances from samples to their respective clusters for k=2,4,6,8,10 respectively. As we increase k, we can see that our performance metric increases thus we should pick k where there is a drastic decrease. In my case I am picking k=4 since there is a significance decrease between 6.3241 and 6.2353.

1.0e+06 *

[6.3241,   6.2353,   6.1279,   6.0598,   6.0207]

```
R=xlsread('nips-87-92.xlsx');
X=R(2:end,3:end);

ks=zeros(1,5);
s=zeros(1,5);
i=1;
for k=2:2:10
[idx,C,sumd]=kmeans(X,k); %elbow method is used, so we pick k at the point
where sumd abrubtly decrease. This is a heuristic algorithm rather than
optimal

    s(i)=sum(sumd);
    ks(i)=k;
    i=i+1;
end
```

Thus, I run the k means algorithm again to Id the documents with their relevant cluster ids k=1,2,3,4

```
R=xlsread('nips-87-92.xlsx');
```

```
X=R(2:end,3:end);
[num,txt,raw]=xlsread('nips-87-92.xlsx','B2:B701');
[idx,C,sumd]=kmeans(X,4); %elbow method is used, so we pick k at the point
where sumd abrubtly decrease. This is a heuristic algorithm rather than
optimal

A=cell(700,1);
B=cell(700,1);
E=cell(700,1);
D=cell(700,1);
for i=1:1:700
    if idx(i)==1
        A(i)=txt(i);

    end
    if idx(i)==2
        B(i)=txt(i);
    end
    if idx(i)==3
        E(i)=txt(i);
    end
    if idx(i)==4
        D(i)=txt(i);
    end

end
```

>> A

A = Cluster 1

'1987_5'
'1989_86'
'1990_112'
'1990_130'
'1991_116'
'1992_3'
'1992_8'
'1992_76'

>> B

B =Cluster 2

'1990_31'
'1990_111'
'1991_21'
'1991_63'
'1991_82'
'1991_86'
'1991_103'
'1991_104'

'1991_112'
'1991_118'
'1991_134'
'1992_2'
'1992_25'
'1992_74'
'1992_90'
'1992_109'

>> E

E = Cluster 3

'1987_2'
'1987_3'
'1987_4'
'1987_8'
'1987_11'
'1987_13'
'1987_14'
'1987_15'
'1987_23'
'1987_24'
'1987_33'
'1987_35'
'1987_36'
'1987_38'
'1987_41'
'1987_44'
'1987_46'
'1987_47'
'1987_50'
'1987_53'
'1987_55'
'1987_56'
'1987_57'
'1987_60'
'1987_62'
'1987_63'
'1987_66'
'1987_67'
'1987_72'
'1987_76'
'1987_77'
'1987_78'
'1987_83'
'1987_84'
'1987_88'

'1987_89'
'1988_1'
'1988_2'
'1988_3'
'1988_4'
'1988_5'
'1988_6'
'1988_9'
'1988_10'
'1988_12'
'1988_13'
'1988_14'
'1988_15'
'1988_16'
'1988_17'
'1988_18'
'1988_19'
'1988_20'
'1988_21'
'1988_24'
'1988_25'
'1988_26'
'1988_27'
'1988_28'
'1988_29'
'1988_30'
'1988_32'
'1988_33'
'1988_36'
'1988_39'
'1988_40'
'1988_42'
'1988_57'
'1988_60'
'1988_61'
'1988_66'
'1988_70'
'1988_75'
'1988_87'
'1988_89'
'1989_13'
'1989_21'
'1989_22'
'1989_23'
'1989_24'
'1989_25'
'1989_26'
'1989_27'
'1989_29'
'1989_30'

'1989_31'
'1989_33'
'1989_39'
'1989_42'
'1989_43'
'1989_45'
'1989_47'
'1989_49'
'1989_50'
'1989_51'
'1989_53'
'1989_54'
'1989_55'
'1989_56'
'1989_57'
'1989_58'
'1989_59'
'1989_60'
'1989_63'
'1989_64'
'1989_65'
'1989_66'
'1989_67'
'1989_68'
'1989_69'
'1989_70'
'1989_73'
'1989_74'
'1989_75'
'1989_76'
'1989_77'
'1989_79'
'1989_80'
'1989_84'
'1989_87'
'1989_90'
'1989_93'
'1989_98'
'1989_101'
'1990_5'
'1990_13'
'1990_20'
'1990_21'
'1990_23'
'1990_26'
'1990_27'
'1990_28'
'1990_29'
'1990_30'
[]

'1990_32'
[]
'1990_34'
[]
'1990_36'
'1990_37'
'1990_40'
'1990_44'
'1990_56'
'1990_58'
'1990_60'
'1990_61'
'1990_68'
'1990_70'

'1990_72'
'1990_73'
'1990_74'
'1990_76'
'1990_77'
'1990_78'
'1990_79'
'1990_82'
'1990_83'
'1990_85'
'1990_86'
'1990_89'
'1990_92'
'1990_93'
'1990_97'
'1990_98'
'1990_99'
'1990_100'
'1990_103'
'1990_104'
'1990_105'
'1990_107'
'1990_108'
'1990_109'
'1990_110'

'1990_116'
'1990_117'

'1990_119'
'1990_121'
'1990_123'
'1990_124'
'1990_126'
'1990_131'

'1990_132'
'1990_137'
'1990_138'

'1991_17'
'1991_18'
'1991_19'
'1991_20'

'1991_22'
'1991_23'
'1991_26'
'1991_29'
'1991_30'
'1991_31'
'1991_32'

'1991_34'

'1991_36'
'1991_37'
'1991_38'
'1991_39'
'1991_40'
'1991_41'
'1991_42'

'1991_47'

'1991_54'

'1991_58'
'1991_60'
'1991_61'
'1991_62'
'1991_67'
'1991_68'

'1991_71'
'1991_74'

'1991_80'
'1991_81'
'1991_84'
'1991_85'

'1991_87'
'1991_89'

'1991_90'
'1991_91'
'1991_96'
'1991_107'
'1991_108'
'1991_110'
'1991_114'
'1991_117'
'1991_119'
'1991_120'
'1991_121'
'1991_122'

'1991_129'
'1991_130'
'1991_132'
'1991_133'
'1991_136'

'1991_138'

'1991_140'

'1991_143'
'1991_144'
'1992_1'

'1992_5'
'1992_6'

'1992_9'
'1992_11'
'1992_12'
'1992_14'
'1992_15'
'1992_17'
'1992_18'
'1992_19'
'1992_20'
'1992_21'
'1992_22'
'1992_23'
'1992_24'

'1992_26'

'1992_29'
'1992_30'
'1992_31'
'1992_35'

'1992_36'

'1992_39'
'1992_42'
'1992_44'
'1992_45'

'1992_49'

'1992_53'
'1992_59'
'1992_61'

'1992_65'

'1992_72'
'1992_75'

'1992_78'
'1992_79'
'1992_80'
'1992_83'
'1992_84'
'1992_85'
'1992_86'
'1992_87'
'1992_88'

'1992_91'

'1992_93'
'1992_103'

'1992_106'
'1992_107'
'1992_115'
'1992_117'

>> D

D =Cluster 4

'1987_1'
'1987_6'
'1987_7'
'1987_9'
'1987_10'
'1987_12'

'1987_16'

'1987_17'
'1987_18'
'1987_19'
'1987_20'
'1987_21'
'1987_22'
'1987_25'
'1987_26'
'1987_27'
'1987_28'
'1987_29'
'1987_30'
'1987_31'
'1987_32'
'1987_34'
'1987_37'
'1987_39'
'1987_40'
'1987_42'
'1987_43'

'1987_45'
'1987_48'
'1987_49'
'1987_51'
'1987_52'
'1987_54'
'1987_58'
'1987_59'
'1987_61'
'1987_64'
'1987_65'
'1987_68'
'1987_69'
'1987_70'
'1987_71'
'1987_73'
'1987_74'
'1987_75'
'1987_79'
'1987_80'
'1987_81'
'1987_82'
'1987_85'
'1987_86'
'1987_87'
'1987_90'
'1988_7'
'1988_8'
'1988_11'

'1988_22'
'1988_23'
'1988_31'

'1988_34'
'1988_35'
'1988_37'
'1988_38'
'1988_41'
'1988_43'
'1988_44'
'1988_45'
'1988_46'
'1988_47'
'1988_48'
'1988_49'
'1988_50'
'1988_51'
'1988_52'
'1988_53'
'1988_54'
'1988_55'
'1988_56'
[]
'1988_58'
'1988_59'
'1988_62'
'1988_63'
'1988_64'
'1988_65'
'1988_67'
'1988_68'
'1988_69'
'1988_71'
'1988_72'
'1988_73'
'1988_74'
'1988_76'
'1988_77'
'1988_78'
'1988_79'
'1988_80'
'1988_81'
'1988_82'
'1988_83'
'1988_84'
'1988_85'
'1988_86'
'1988_88'
'1988_90'

'1988_91'
'1988_92'
'1988_93'
'1988_94'
'1988_95'
'1989_1'
'1989_2'
'1989_3'
'1989_4'
'1989_5'
'1989_6'
'1989_7'
'1989_8'
'1989_9'
'1989_10'
'1989_11'
'1989_12'
'1989_14'
'1989_15'
'1989_16'
'1989_17'
'1989_18'
'1989_19'
'1989_20'
'1989_28'
'1989_32'
'1989_34'
'1989_35'
'1989_36'
'1989_37'
'1989_38'
'1989_40'
'1989_41'
'1989_44'
'1989_46'
'1989_48'
'1989_52'
'1989_61'
'1989_62'
'1989_71'
'1989_72'
'1989_78'
'1989_81'
'1989_82'
'1989_83'
'1989_85'

'1989_88'
'1989_89'
'1989_91'

'1989_92'
'1989_94'
'1989_95'
'1989_96'
'1989_97'
'1989_99'
'1989_100'
'1990_1'
'1990_2'
'1990_3'
'1990_4'
'1990_6'
'1990_7'
'1990_8'
'1990_9'
'1990_10'
'1990_11'
'1990_12'
'1990_14'
'1990_15'
'1990_16'
'1990_17'
'1990_18'
'1990_19'
'1990_22'
'1990_24'
'1990_25'
'1990_33'
'1990_35'
'1990_38'
'1990_39'
'1990_41'
'1990_42'
'1990_43'
'1990_45'
'1990_46'
'1990_47'
'1990_48'
'1990_49'
'1990_50'
'1990_51'
'1990_52'
'1990_53'
'1990_54'
'1990_55'
'1990_57'
'1990_59'
'1990_62'
'1990_63'
'1990_64'

'1990_65'
'1990_66'
'1990_67'
'1990_69'

'1990_71'
'1990_75'
'1990_80'
'1990_81'

'1990_84'
'1990_87'
'1990_88'
'1990_90'
'1990_91'
'1990_94'
'1990_95'
'1990_96'

'1990_101'
'1990_102'
'1990_106'
'1990_113'
'1990_114'
'1990_115'
'1990_118'
'1990_120'

'1990_122'

'1990_125'
'1990_127'
'1990_128'
'1990_129'
'1990_133'
'1990_134'
'1990_135'
'1990_136'
'1990_139'
'1990_140'
'1990_141'
'1990_142'
'1990_143'
'1991_1'
'1991_2'
'1991_3'
'1991_4'
'1991_5'
'1991_6'
'1991_7'

'1991_8'
'1991_9'
'1991_10'
'1991_11'
'1991_12'
'1991_13'
'1991_14'
'1991_15'
'1991_16'
'1991_24'
'1991_25'
'1991_27'
'1991_28'
'1991_33'
'1991_35'
'1991_43'
'1991_44'
'1991_45'
'1991_46'
'1991_48'
'1991_49'
'1991_50'
'1991_51'
'1991_52'
'1991_53'
'1991_55'
'1991_56'
'1991_57'
'1991_59'
'1991_64'
'1991_65'
'1991_66'
'1991_69'
'1991_70'
'1991_72'
'1991_73'
'1991_75'
'1991_76'
'1991_77'
'1991_78'
'1991_79'
'1991_83'
'1991_88'
'1991_92'
'1991_93'
'1991_94'
'1991_95'
'1991_97'
'1991_98'
'1991_99'

'1991_100'
'1991_101'
'1991_102'
'1991_105'
'1991_106'
'1991_109'
'1991_111'
'1991_113'
'1991_115'
'1991_123'
'1991_124'
'1991_125'
'1991_126'
'1991_127'
'1991_128'
'1991_131'
'1991_135'
'1991_137'
'1991_139'
'1991_141'
'1991_142'
'1992_4'

'1992_7'
'1992_10'
'1992_13'
'1992_16'
'1992_27'
'1992_28'
'1992_32'
'1992_33'
'1992_34'
'1992_37'
'1992_38'

'1992_40'
'1992_41'
'1992_43'
'1992_46'
'1992_47'
'1992_48'

'1992_50'
'1992_51'
'1992_52'
'1992_54'
'1992_55'
'1992_56'
'1992_57'
'1992_58'

'1992_60'
'1992_62'
'1992_63'
'1992_64'
'1992_66'
'1992_67'
'1992_68'
'1992_69'
'1992_70'
'1992_71'
'1992_73'
'1992_77'
'1992_81'
'1992_82'
'1992_89'
'1992_92'
'1992_94'
'1992_95'
'1992_96'
'1992_97'
'1992_98'
'1992_99'
'1992_100'
'1992_101'
'1992_102'
'1992_104'
'1992_105'
'1992_108'
'1992_110'
'1992_111'
'1992_112'
'1992_113'
'1992_114'
'1992_116'
'1992_118'
'1992_119'
'1992_120'
'1992_121'
'1992_122'
'1992_123'
'1992_124'
'1992_125'
'1992_126'
'1992_127'