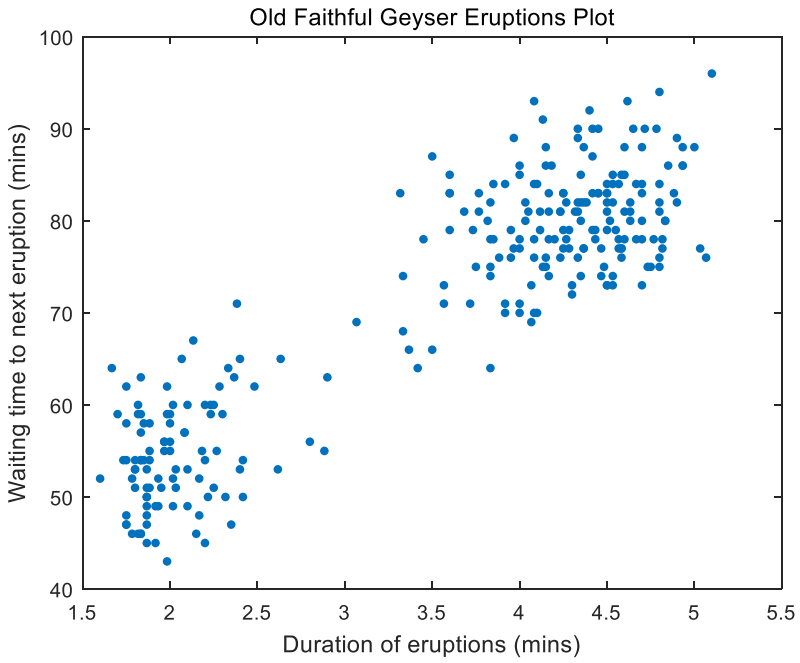
# EE511 Project 3

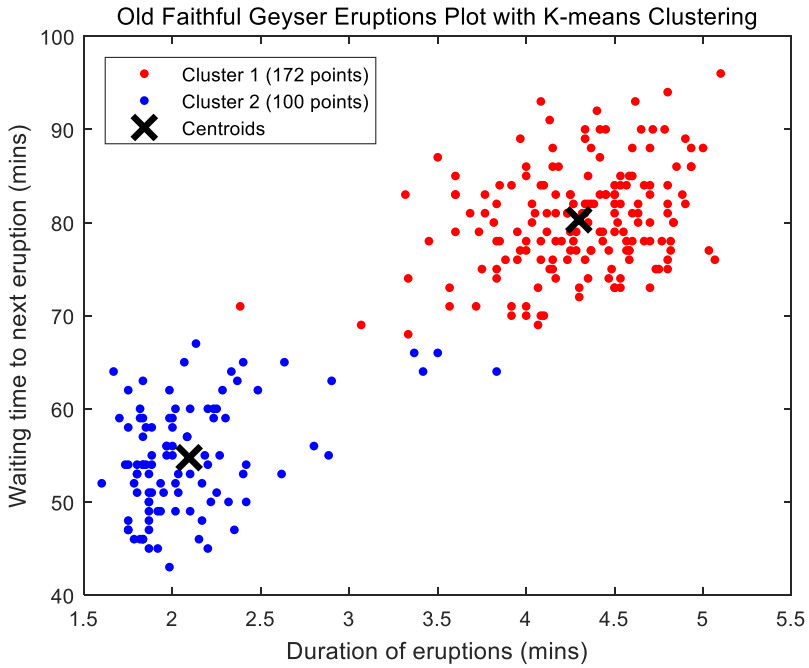
This project is implemented using MATLAB. Codes are attached in the end of this document and also available at https://github.com/uscwy/ee511project3/blob/master/project3.m

**[Testing Faith]**

In this problem, I use textscan function to read data from text files and convert to a 272 by 2 matrix. The first column is duration of eruption and the second column is the waiting time to next eruption. Then generate the scatter plot as following.




Then use k-mean clustering function to partition the data into two clusters (k = 2). This k-means function uses the squared Euclidean distance measure and the k-means++ algorithm for cluster center initialization. The number of points in cluster and centroids keep the same in each run.

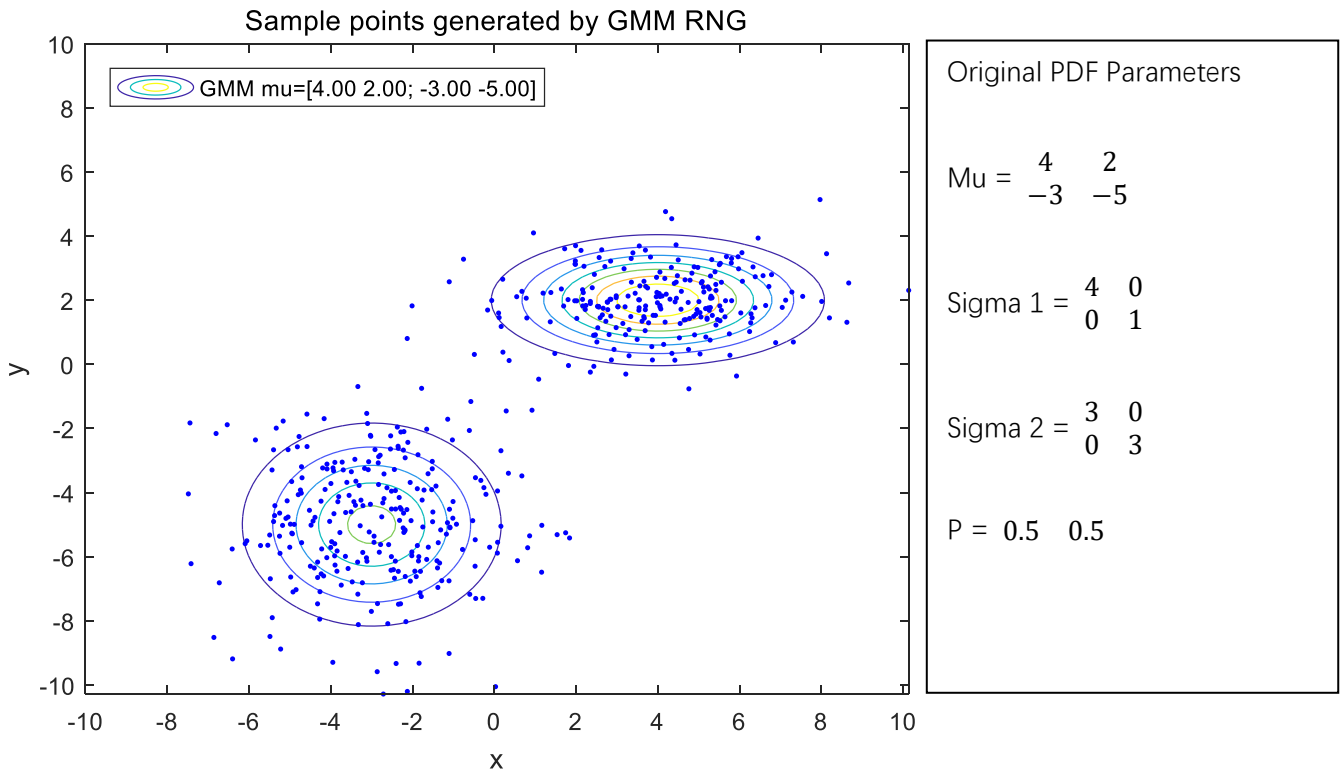


**[EM]**

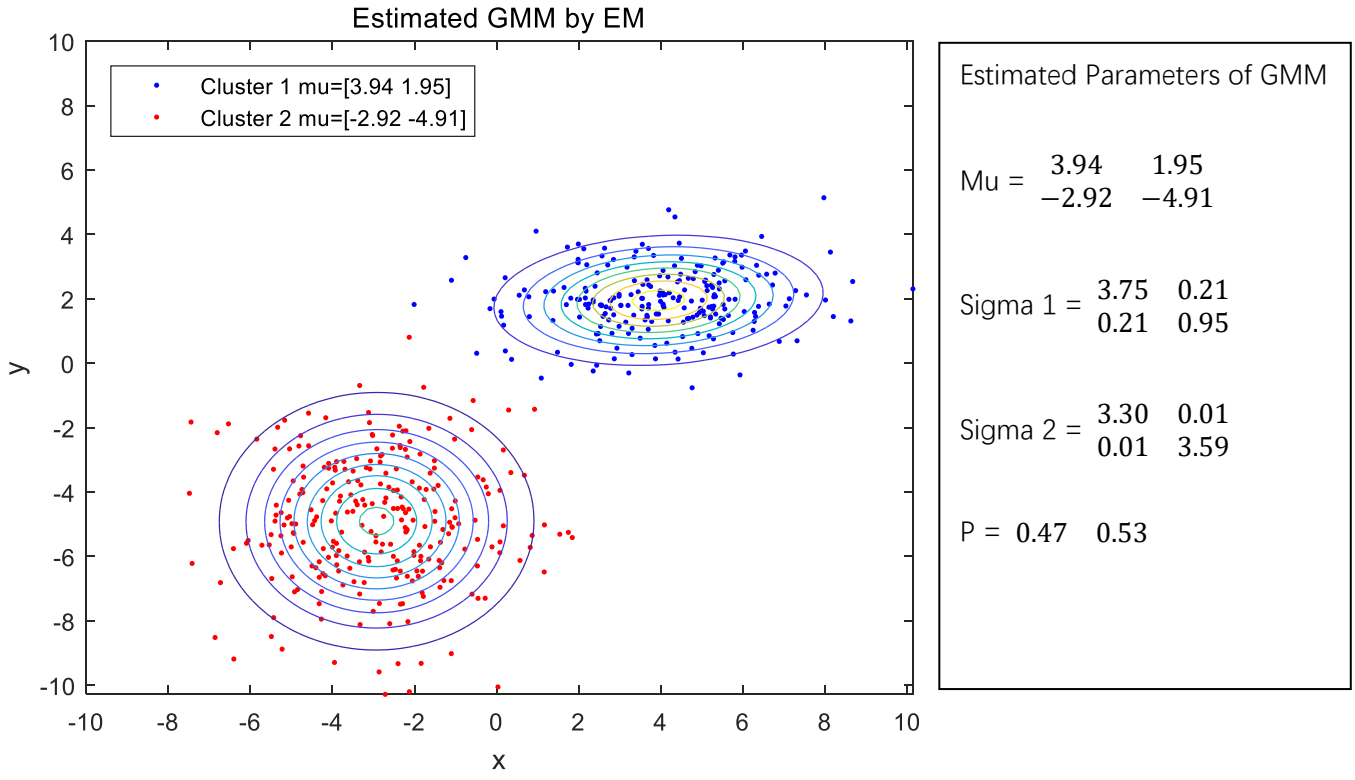
GMM RNG with 2 subpopulations:

The below plot is the contour of pdf and sampling points generated by this RNG.

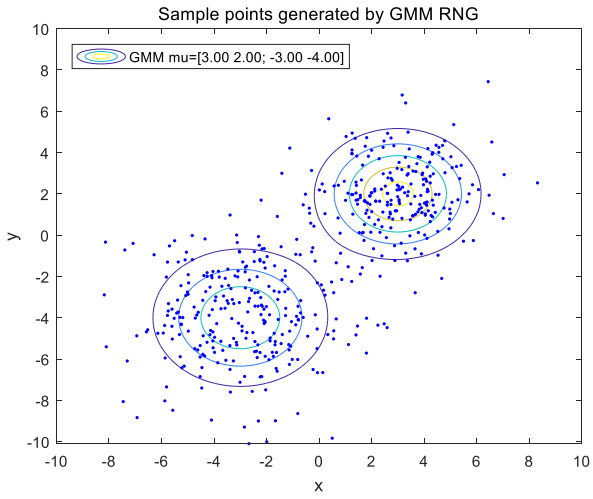
The original parameters of pdf used by this RNG are listed on right side.

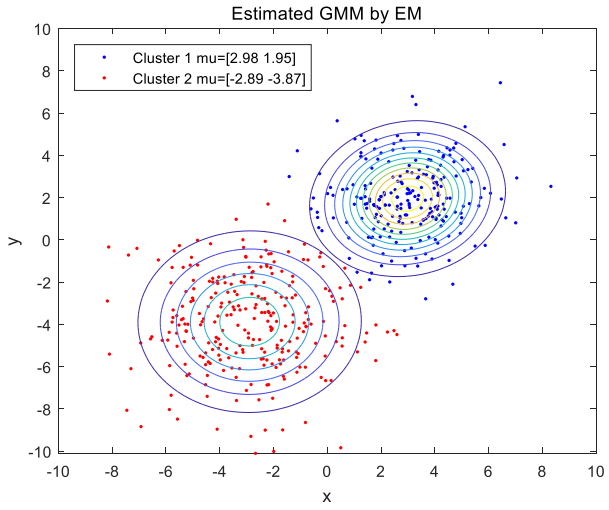


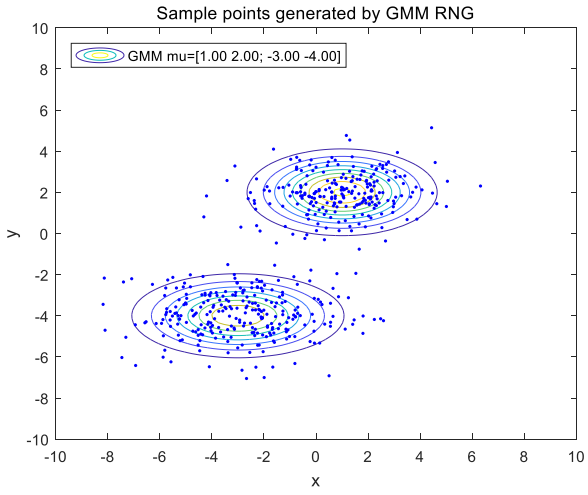

Then use expectation maximization (EM) algorithm to estimate the pdf of sampling points. The below plot is the contour of estimated pdf. The estimated parameters are listed on right side. Comparing with original parameters and the shape of pdf, the parameters and shape conform to the original approximately.

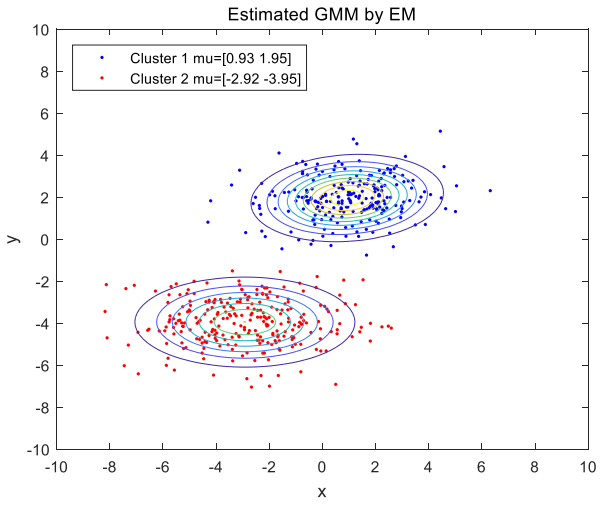


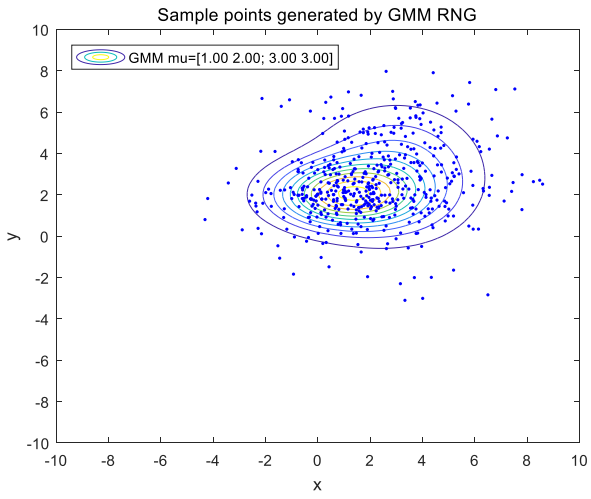
The following plots are estimated pdf of spherical covariance matrices, ellipsoidal covariance matrices and poorly-separated subpopulations. Left side are original pdf and right side are estimated pdf. For the quality, spherical and ellipsoidal covariance are better than poorly-separated. For speed of estimation, spherical and ellipsoidal covariance are much faster than poorly-separated. Spherical and ellipsoidal cases generally take dozens of iterations, but poorly-separated case may take hundreds of iterations.

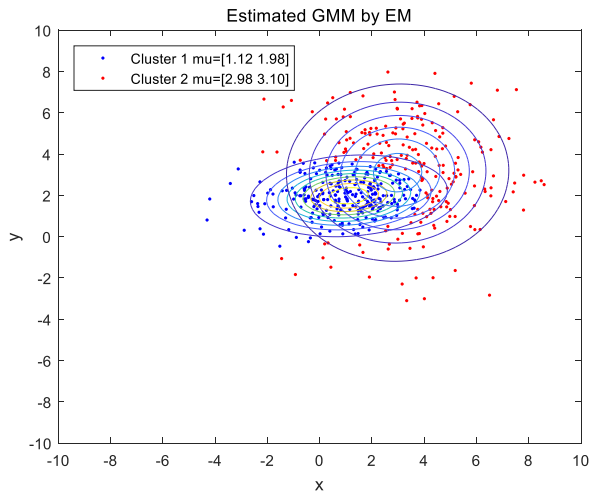


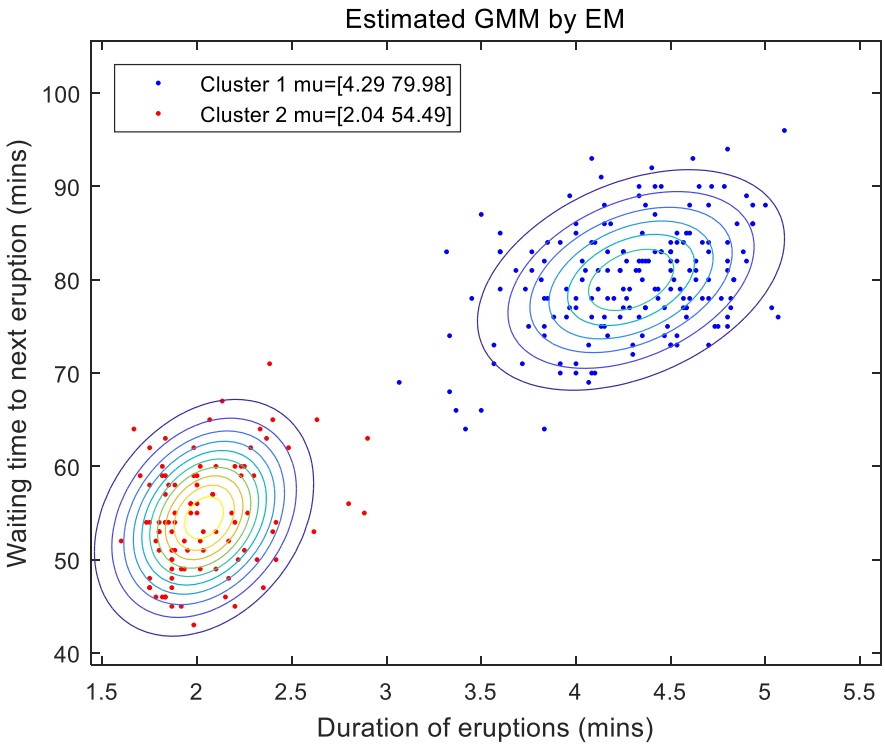



The below plot is the GMM-EM estimation for old faithful data set.




Estimated GMM for Old faithful data

$$\text{Mu} = \begin{matrix} 4.29 & 79.88 \\ 2.04 & 54.49 \end{matrix}$$

$$\text{Sigma 1} = \begin{matrix} 0.17 & 0.93 \\ 0.93 & 35.88 \end{matrix}$$

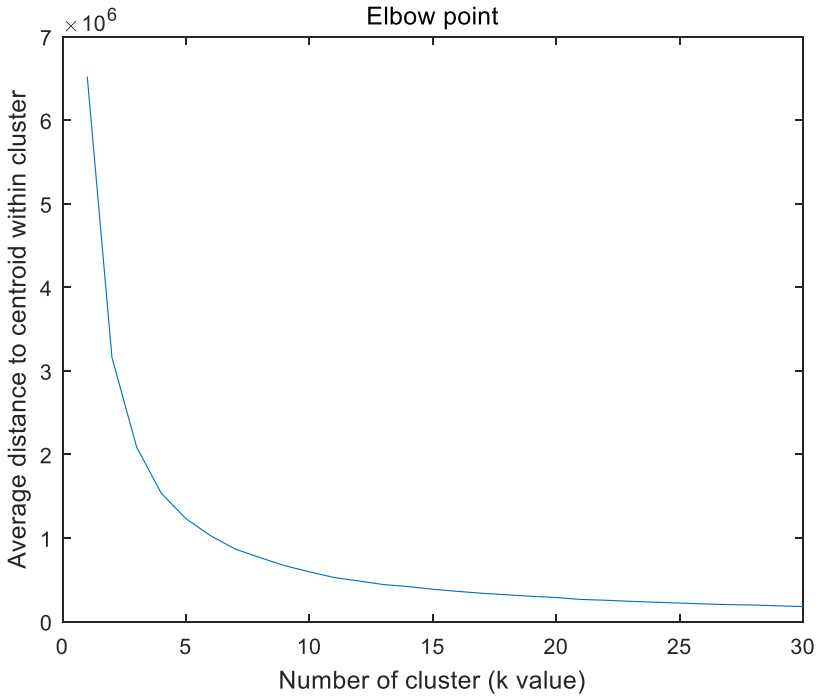
$$\text{Sigma 2} = \begin{matrix} 0.07 & 0.44 \\ 0.44 & 33.77 \end{matrix}$$

$$\text{P} = \begin{matrix} 0.64 & 0.36 \end{matrix}$$

**[Clusters of Text]**

I tried different k values and record the average distance to centroid within cluster. The below plot it the relationship between k value and average distance to centroid within cluster. I choose k = 12 because the average distance changes little when k goes greater than 12.

Then I use k = 12 to run k-means 10 times and calculate the variance of document number in each cluster to find the smallest variance as the result.




Cluster 1 (158 docs)

---

1987_7 1987_9 1987_10 1987_12 1987_16 1987_17 1987_20 1987_21 1987_26 1987_29
1987_31 1987_32 1987_39 1987_43 1987_48 1987_49 1987_51 1987_59 1987_69 1987_71
1987_73 1987_74 1987_82 1987_86 1988_22 1988_34 1988_41 1988_43 1988_44 1988_45
1988_46 1988_47 1988_48 1988_49 1988_50 1988_51 1988_52 1988_54 1988_55 1988_56
1988_65 1988_67 1988_71 1988_77 1988_78 1988_79 1988_81 1988_82 1988_83 1988_84
1988_91 1988_95 1989_1 1989_2 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_35 1989_36 1989_38 1989_41 1989_91 1989_96 1990_1 1990_2 1990_3 1990_4
1990_6 1990_7 1990_8 1990_12 1990_14 1990_17 1990_18 1990_19 1990_22 1990_38
1990_43 1990_46 1990_47 1990_49 1990_51 1990_52 1990_53 1990_55 1990_69 1990_84
1991_1 1991_2 1991_4 1991_5 1991_6 1991_7 1991_8 1991_9 1991_10 1991_11
1991_12 1991_13 1991_15 1991_16 1991_24 1991_35 1991_43 1991_44 1991_51 1991_52
1991_66 1991_73 1991_75 1991_76 1991_77 1991_88 1991_93 1991_94 1991_100 1992_16
1992_46 1992_47 1992_51 1992_52 1992_54 1992_64 1992_69 1992_71 1992_82 1992_96
1992_100 1992_101 1992_102 1992_105 1992_108 1992_111 1992_112 1992_113 1992_114 1992_116
1992_120 1992_121 1992_122 1992_123 1992_124 1992_125 1992_126 1992_127

Cluster 2 (7 docs)

---

1987_47 1987_53 1987_66 1989_50 1991_61 1991_62 1992_55

Cluster 3 (52 docs)
---
1987_2 1987_4 1987_27 1987_57 1987_67 1988_5 1988_23 1988_87 1989_40 1989_58
1989_67 1989_84 1989_87 1989_90 1990_59 1990_62 1990_63 1990_64 1990_65 1990_68
1990_89 1990_109 1990_138 1991_14 1991_31 1991_36 1991_41 1991_64 1991_65 1991_69
1991_70 1991_107 1991_144 1992_9 1992_10 1992_11 1992_20 1992_31 1992_33 1992_34
1992_36 1992_37 1992_38 1992_40 1992_41 1992_42 1992_63 1992_78 1992_103 1992_115
1992_118 1992_119

Cluster 4 (56 docs)
---
1987_65 1988_4 1988_7 1988_30 1988_35 1988_68 1988_72 1989_37 1989_48 1989_56
1989_66 1989_75 1989_83 1990_25 1990_50 1990_61 1990_73 1990_90 1990_94 1990_96
1990_102 1990_108 1990_111 1990_124 1990_131 1990_134 1991_45 1991_48 1991_50 1991_63
1991_72 1991_79 1991_81 1991_82 1991_86 1991_103 1991_110 1991_118 1991_131 1991_134
1991_135 1991_137 1992_2 1992_7 1992_25 1992_32 1992_44 1992_50 1992_66 1992_67
1992_72 1992_74 1992_90 1992_91 1992_92 1992_109

Cluster 5 (82 docs)
---
1987_8 1987_13 1987_38 1987_44 1987_63 1987_72 1987_76 1987_83 1987_88 1988_1
1988_2 1988_3 1988_9 1988_12 1988_13 1988_14 1988_17 1988_18 1988_19 1988_24
1988_25 1988_27 1988_36 1988_60 1988_70 1988_75 1989_22 1989_39 1989_45 1989_53
1989_54 1989_55 1989_59 1989_63 1989_64 1989_79 1989_98 1990_5 1990_13 1990_20
1990_21 1990_23 1990_26 1990_28 1990_58 1990_70 1990_76 1990_77 1990_86 1990_92
1990_93 1990_97 1990_104 1990_105 1990_123 1991_29 1991_32 1991_34 1991_40 1991_42
1991_67 1991_68 1991_71 1991_74 1991_87 1991_112 1991_119 1991_121 1991_133 1991_138
1992_5 1992_12 1992_15 1992_17 1992_21 1992_22 1992_30 1992_35 1992_45 1992_65
1992_106 1992_107

Cluster 6 (1 docs)
---
1992_83

Cluster 7 (47 docs)
---
1987_41 1988_16 1988_57 1988_58 1989_21 1989_23 1989_24 1989_26 1989_68 1989_69
1989_70 1989_73 1989_76 1989_77 1990_29 1990_31 1990_119 1990_121 1990_126 1990_132
1991_17 1991_19 1991_20 1991_21 1991_37 1991_38 1991_58 1991_84 1991_104 1991_120
1991_122 1991_129 1991_132 1991_136 1991_140 1991_143 1992_6 1992_18 1992_19 1992_26
1992_29 1992_59 1992_61 1992_80 1992_86 1992_87 1992_88

Cluster 8 (2 docs)
---
1987_55 1988_66

Cluster 9 (91 docs)
---

1987_15 1987_40 1987_78 1987_87 1988_15 1988_26 1988_28 1988_29 1988_32 1988_33
1988_37 1988_38 1988_39 1988_40 1988_42 1988_63 1988_85 1988_94 1989_13 1989_25
1989_27 1989_29 1989_30 1989_31 1989_32 1989_33 1989_34 1989_49 1989_51 1989_52
1989_100 1990_27 1990_30 1990_32 1990_33 1990_34 1990_35 1990_36 1990_37 1990_39
1990_40 1990_41 1990_42 1990_48 1990_56 1990_60 1990_67 1990_74 1990_78 1990_79
1990_82 1990_83 1990_107 1990_110 1990_116 1990_141 1990_142 1991_18 1991_22 1991_23
1991_25 1991_26 1991_30 1991_46 1991_47 1991_54 1991_55 1991_56 1991_57 1991_59
1991_60 1991_78 1991_80 1991_90 1991_91 1991_96 1991_99 1991_101 1992_4 1992_14
1992_24 1992_39 1992_49 1992_53 1992_81 1992_84 1992_85 1992_89 1992_93 1992_98
1992_117

Cluster 10 (3 docs)
---
1987_35 1987_75 1989_3

Cluster 11 (17 docs)
---
1987_11 1987_14 1987_46 1987_56 1987_60 1987_62 1987_77 1987_84 1988_6 1988_89
1989_47 1989_86 1989_93 1989_101 1991_39 1991_89 1992_23

Cluster 12 (183 docs)
---
1987_3 1987_5 1987_6 1987_18 1987_19 1987_22 1987_23 1987_24 1987_25 1987_28
1987_30 1987_33 1987_34 1987_36 1987_37 1987_42 1987_45 1987_50 1987_52 1987_54
1987_58 1987_61 1987_64 1987_68 1987_70 1987_79 1987_80 1987_81 1987_85 1987_89
1987_90 1988_8 1988_10 1988_11 1988_20 1988_21 1988_31 1988_53 1988_59 1988_61
1988_62 1988_64 1988_69 1988_73 1988_74 1988_76 1988_80 1988_86 1988_88 1988_90
1988_92 1988_93 1989_28 1989_42 1989_43 1989_44 1989_46 1989_57 1989_60 1989_61
1989_62 1989_65 1989_71 1989_72 1989_74 1989_78 1989_80 1989_81 1989_82 1989_85
1989_88 1989_89 1989_92 1989_94 1989_95 1989_97 1989_99 1990_9 1990_10 1990_11
1990_15 1990_16 1990_24 1990_44 1990_45 1990_54 1990_57 1990_66 1990_71 1990_72
1990_75 1990_80 1990_81 1990_85 1990_87 1990_88 1990_91 1990_95 1990_98 1990_99
1990_100 1990_101 1990_103 1990_106 1990_112 1990_113 1990_114 1990_115 1990_117 1990_118
1990_120 1990_122 1990_125 1990_127 1990_128 1990_129 1990_130 1990_133 1990_135 1990_136
1990_137 1990_139 1990_140 1990_143 1991_3 1991_27 1991_28 1991_33 1991_49 1991_53
1991_83 1991_85 1991_92 1991_95 1991_97 1991_98 1991_102 1991_105 1991_106 1991_108
1991_109 1991_111 1991_113 1991_114 1991_115 1991_116 1991_117 1991_123 1991_124 1991_125
1991_126 1991_127 1991_128 1991_130 1991_139 1991_141 1991_142 1992_1 1992_3 1992_8
1992_13 1992_27 1992_28 1992_43 1992_48 1992_56 1992_57 1992_58 1992_60 1992_62
1992_68 1992_70 1992_73 1992_75 1992_76 1992_77 1992_79 1992_94 1992_95 1992_97
1992_99 1992_104 1992_110

```
%EE511 Project 3
%Author: Yong Wang <yongw@usc.edu>

%% [Testing Faith]
f = fopen('old-faithful.txt');
d = textscan(f, '%d %f %f', 'HeaderLines', 26);
fclose(f);
duration = d{1,2};
waiting = d{1,3};
%scatter(duration,waiting);
plot(duration,waiting,'.', 'MarkerSize',12);
title('Old Faithful Geyser Eruptions Plot');
xlabel('Duration of eruptions (mins)');
ylabel('Waiting time to next eruption (mins)');
%k-means clustering, k=2%
X = [duration, waiting];
[idx,C] = kmeans(X,2);
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)
title('Old Faithful Geyser Eruptions Plot with K-means Clustering');
xlabel('Duration of eruptions (mins)');
ylabel('Waiting time to next eruption (mins)');
legend(sprintf('Cluster 1 (%d points)',length(X(idx==1,1))),...
      sprintf('Cluster 2 (%d points)',length(X(idx==2,1))),...
      'Centroids','Location','NW');
hold off

%% [EM]
mu_RNG = [4 2;-3 -5];
sigma_RNG = cat(3,[4 0;0 1],[3 0;0 3]);
p_RNG = ones(1,2)/2;
%mu_RNG = [3 2;-3 -4];
%sigma_RNG = cat(3,[3 0;0 3],[4 0;0 4]);
%p_RNG = ones(1,2)/2;

gm = gmdistribution(mu_RNG,sigma_RNG,p_RNG);
gmPDF = @(x,y)pdf(gm,[x y]);
rng(1);
%Use GMM RNG to generate sample points
s = 500; %number of sampling points
Y = random(gm,s);



figure;
axis([-10 10 -10 10]);
li = linspace(-10, 10, 100);
li2 = linspace(-10, 10, 100);
```

```
[A B] = meshgrid(li, li);
grid = [A(:), B(:)];


z1 = pdf(gm, grid);
z1 = reshape(z1, 100, 100);
contour(li, li2, z1);
hold on
scatter(Y(:,1),Y(:,2), 'b.')
title('Sample points generated by GMM RNG');
legend(sprintf('GMM mu=[%1.2f %1.2f; %1.2f %1.2f]',mu_RNG(1,1),mu_RNG(1,2),...
      mu_RNG(2,1),mu_RNG(2,2)),...
      'Location','NW');
xlabel('x');
ylabel('y');
hold off

%use EM GMM for old faithful
% Y = X; s = length(X);
% axis auto;
% li = linspace(min(Y(:,1))*0.9, max(Y(:,1))*1.1, 100);
% li2 = linspace(min(Y(:,2))*0.9, max(Y(:,2))*1.1, 100);
% [A B] = meshgrid(li, li2);
% grid = [A(:), B(:)];

%Use ME to evaluate parameter for sample points Y
k = 2;  % The number of subpopulation.

%randomly select k points as inital mean
i = randi(s, k, 1);
mu = Y(i(1:k), :);
%initial sigma
sigma = [];
for i = 1 : k
    sigma{i} = cov(Y);
end
%inital p
p = ones(1, k)/k;
%record which cluster the data points belong to
idx = zeros(s, k);
P = zeros(s, k);
%search for EM
for i = 1 : 3000

    Z = zeros(s, k); %record value from pdf
    for j = 1 : k
        gm = gmdistribution(mu(j, :), sigma{j}, p(j));
        % Evaluate the Gaussian for all data points for cluster 'j'.
        Z(:,j) = pdf(gm,Y);
```

```
    end
    P = bsxfun(@rdivide, Z, sum(Z, 2));

    %start maximization
    pmu = mu;
    for j=1:k %for each cluster
        p(j) = mean(P(:,j), 1);
        mu(j,:) = (P(:, j)' * Y) ./ sum(P(:,j), 1);
        %calculate covariance
        Ym = bsxfun(@minus, Y, mu(j, :));
        sigma_k = zeros(2);
        for t = 1 : s
            sigma_k = sigma_k + (P(t, j) .* (Ym(t, :)' * Ym(t, :)));
        end
        sigma{j} = sigma_k ./ sum(P(:, j));
    end
    fprintf('%d: mu=[%1.2f %1.2f;%1.2f %1.2f] p=[%1.2f %1.2f]\n',...
           i, mu(1,1),mu(1,2),mu(2,1),mu(2,2),p(1),p(2));
    if(mu == pmu)
        break;
    end
end
%scatter plot
figure;

gm = gmdistribution(mu(1, :), sigma{1}, p(1));
z1 = pdf(gm, grid);
gm = gmdistribution(mu(2, :), sigma{2}, p(2));
z2 = pdf(gm, grid);
z1 = reshape(z1, 100, 100);
z2 = reshape(z2, 100, 100);
cluster1 = (find(P(:,1)>0.5));
cluster2 = (find(P(:,2)>0.5));
plot(Y(cluster1(:),1), Y(cluster1(:),2), 'b.');
hold on
plot(Y(cluster2(:),1), Y(cluster2(:),2), 'r.');
hold on
contour(li, li2, z1);
contour(li, li2, z2);
legend(sprintf('Cluster 1 mu=[%1.2f %1.2f]',mu(1,1),mu(1,2)),...
      sprintf('Cluster 2 mu=[%1.2f %1.2f]',mu(2,1),mu(2,2)),...
      'Location','NW');
title('Estimated GMM by EM');
xlabel('x');
ylabel('y');
%% [Clusters of Text]

%T = readtable('nips-87-92.csv');
[r, c] = size(T);
```

```
fid = T{2:r,2};
D = T{2:r, 3:c};
k = 12;
avd = zeros(k,1);
for i=k:k
    [idx, centroid, sumd] = kmeans(D,i);
    avd(i) = sum(sumd)/i;
end

for i=1:k
    doc = find(idx==i);
    fprintf("\n\nCluster %d (%d docs)\n---------\n", i, length(doc));
    for j=1:length(doc)
        fprintf("%s ", fid{doc(j)});
        if(mod(j,10) == 0)
            fprintf("\n");
        end
    end
end
figure;
plot([1:k], avd);
% silhouette(D,idx);
xlabel("Number of cluster (k value)");
ylabel('Average distance to centroid within cluster');
title('Elbow point');

% legend(sprintf('k=%d', k),'Location','NE');

% f = fopen('nips-87-92.csv');
% t = readtable('nips-87-92.csv');
% headline = fgetl(f);
% words = textscan(headline, "%s",'Delimiter',',','EmptyValue',-Inf);
% while ~feof(f)
%     l = fgetl(f);
%     [fid, pos] = textscan(l,"%d,%s,", 'Delimiter', ',');
%     %fprintf("fid=%d\n",fid{1,1});
%     c = textscan(l(pos+1:end), "%d", 'Delimiter', ',');
% end
% fclose(f);
```