UNIVERSITY OF BIRMINGHAM

Machine Learning (Extended)

Computer Based Test 3

Clustering

December 4, 2016

Thomas Brereton 1708846

Table of Content

List of Figures

List of Tables

1 Overall Task

In this computer based test we are given a data file containing functional magnetic resonance imaging (fMRI) results. This data set consists of fMRI scans of 90 brain (cortical) regions over 200 time points, for 40 subjects (20 healthy and 20 diseased). We are tasked with clustering the similar regions according there functioning over time i.e. time points are the dimensions/features. Additionally, we also vary the number of clusters, K, and comment on its effect on the grouping of similar brain regions. We support our results with a 'community matrix [1],' which helps us visualise the connected brain regions of the brain. In summary the tasks for this report are outlined as per the following.

- Perform k-means clustering on this data to group brain regions into K=10 clusters
- Perform k-means clustering by varying the number of clusters (K=20, and K=30)
- Generate community matrix for K=10, K=20, and K=30 for healthy as well as for diseased groups.
- Compare and comment on the results by looking at the differences between community matrices of healthy and diseased groups and the effect of K.

The analysis was done in Matlab and the code listing for the clustering is found in Appendix A. Additionally, the author looked at optimising the number of clusters by producing an 'elbow plot,' the code listing for this can be found in Appendix B.

This report will first provide comments on the results for 10, 20, and 30 clusters. Following these sections is the main body of the report, which will focus on the comparison between the healthy and diseased community matrices and the effects of varying the number of clusters.

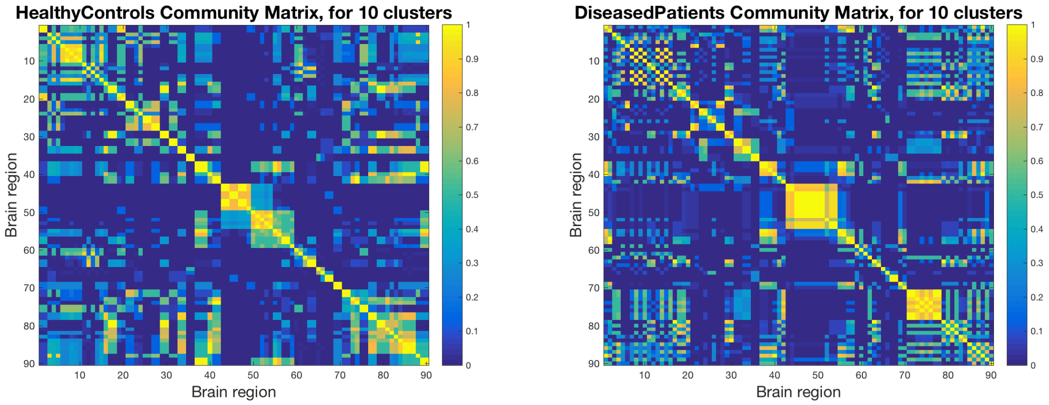
2 Results for 10 Clusters

In this section we present the results from clustering the brain regions across 90 time points for K = 10. The results are in the form of community matrices, which visualises the "coherence in neuro-activities across [the] two hemispheres" [1]. The community matrix is achieved by comparing the ith brain region cluster index with all 90 other brain regions cluster indexes and return a 1 if they are identical and 0 if not. This is then repeated for every brain region to produce a 90 by 90 matrix. Furthermore, we take the community matrices averaged across the 20 respective patients to ensure consistent results.

In the community matrices, the brain regions with strong coherence in neuro-activities across the two hemispheres (strong connectivity) are represented with a yellow (1 on the colour scale), and with weak connectivity are represented with a dark blue (0 on the colour scale) . Therefore, the negative diagonal with value 1 is expected as it is the comparison of the same brain regions.

Looking at the healthy controls in Figure 1a we can see strong connectivity surrounding the diagonal in the top-left and bottom-right corners. This indicates "strong coherence in neuro-activities across two hemi- spheres" [1]. The diseased patients in Figure 1b show less occurrences of strong connectivity surround the diagonal, however, there is two large blocks in the middle and bottom-right showing stronger connectivity across more brain regions. We also see medium strong connections in far off diagonal elements in both Figures 1a and 1b.

The 'Difference Community Matrix' is the combined difference of the healthy controls and diseased patients. A yellow, or high value, indicates a strong connectivity in healthy controls, a blue, or low value, indicates a strong connectivity in diseased patients, and a green, or zero value, indicates there is equal connectivity between healthy and diseased patients.



(a) Healthy controls community matrix for 10 clusters (b) Diseased patients community matrix for 10 clusters

Figure 1: Community matrices for 10 clusters

The difference matrix in Figure 2 illustrates that there is generally a stronger connectivity in healthy controls rather than in diseased patients.

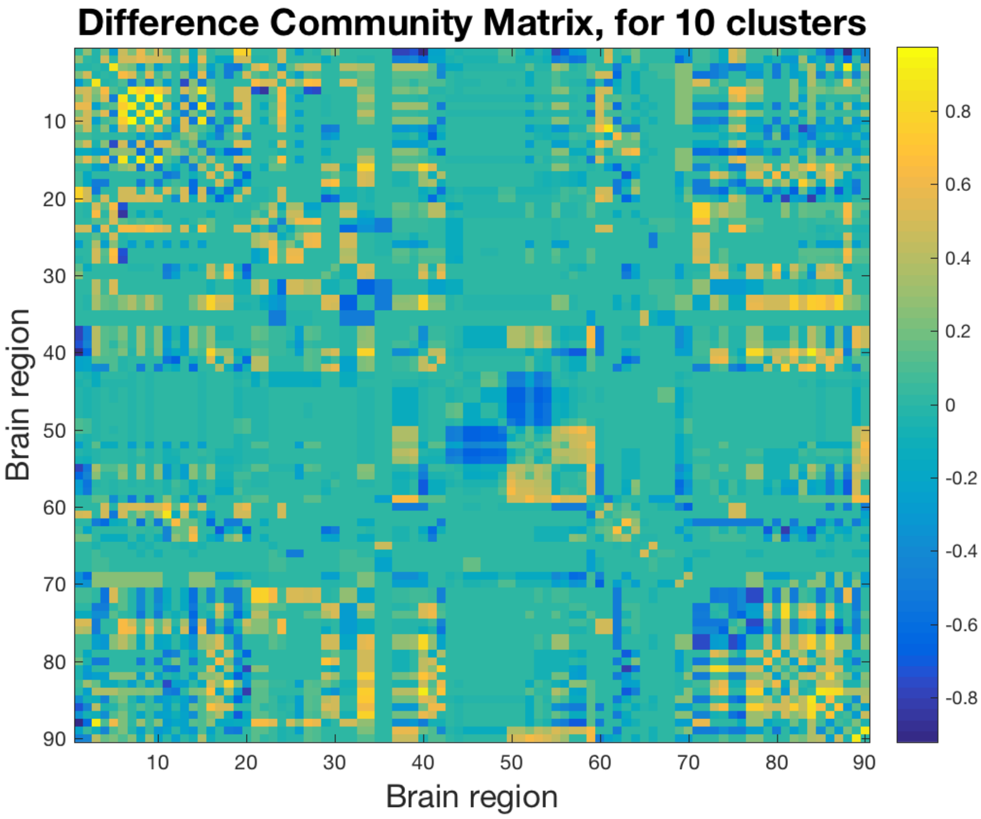
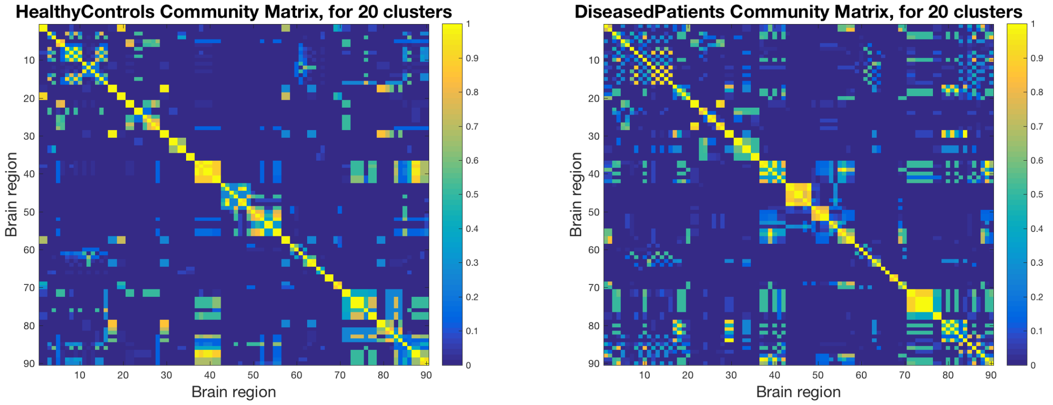


Figure 2: Difference community matrix for 10 clusters

3 Results for 20 Clusters

In this section we look at the same community matrices as described previously, but for a total of 20 clusters.

Looking at the healthy controls in Figure 3a we can see strong connectivity consistently surrounding the diagonal. The diseased patients in Figure 3b show similar results to before with the two same large blocks in the middle and bottom-right showing stronger connectivity across more brain regions. Interestingly, with 20 clusters less connectivity is shown in the far off diagonal brain regions in both Figures 3a and 3b.



(a) Healthy controls community matrix for 20 clusters (b) Diseased patients community matrix for 20 clusters

Figure 3: Community matrices for 20 clusters

The difference matrix in Figure 4 illustrates that there is generally a stronger connectivity in healthy controls rather than in diseased patients but there are more strong connections for less connectivity in far off diagonal brain regions.

4 Results for 30 Clusters

In this section we look at the same community matrices as described previously, but for a total of 30 clusters.

Performing k-means with 30 clusters yield similar results to 20 clusters, except possibly more refined. Looking at the healthy controls in Figure 5a we can see vibrant yellows surround the diagonal as highlighted by the red boxes. The diseased patients in Figure 5b show less strong connectivity occurring near the as shown by the middle two red boxes. The outer two boxes show that there is a decrease in the strength of connectivity but more connections with surrounding regions are made. As for 20 clusters, 30 clusters shows even less connectivity is shown in the far off diagonal brain regions as illustrated in both Figures 5a and 5b.

The difference matrix in Figure 6 shows that there is still stronger connectivity in the healthy controls, however, there are occurrences of distinct, strong, connections in diseased patients as highlighted by the red boxes.

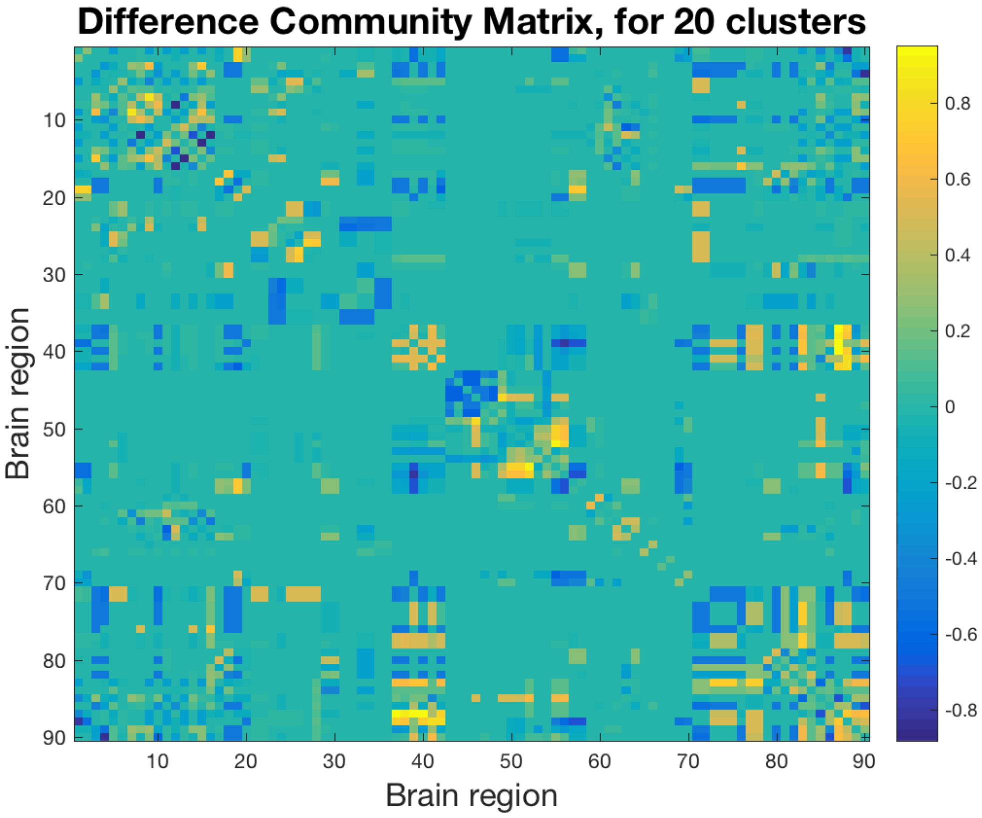
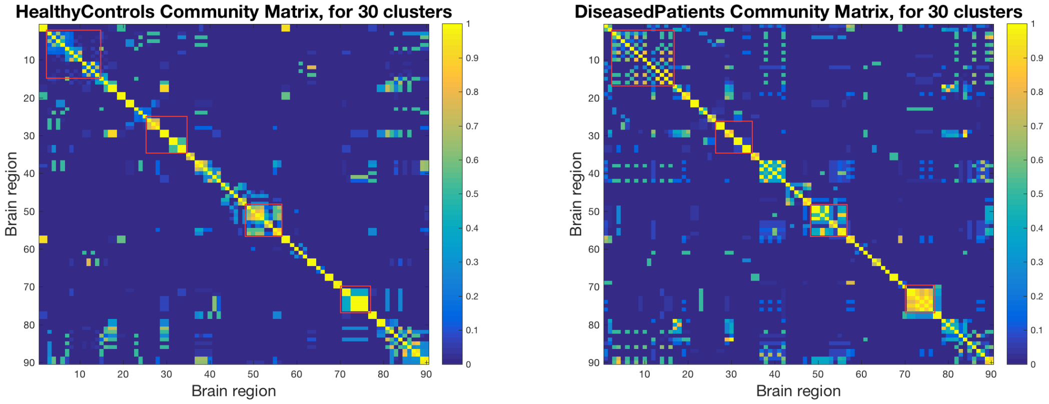


Figure 4: Difference community matrix for 20 clusters



(a) Healthy controls community matrix for 30 clusters (b) Diseased patients community matrix for 30 clusters

Figure 5: Community matrices for 30 clusters

5 Comparison of number of clusters

When comparing the results for the varying number of clusters we see a trend that there is less connectivity among the brain regions. This is evidenced by looking at the increase of 'blue' when looking at Figure 1 to 5. In turn, this means only the brain regions with strong connectivity are highlighted.

This occurrence can be explained by the number of clusters. With a lower number of clusters, there is more chance one brain region is in the same cluster as another. Conversely, there is a smaller chance of

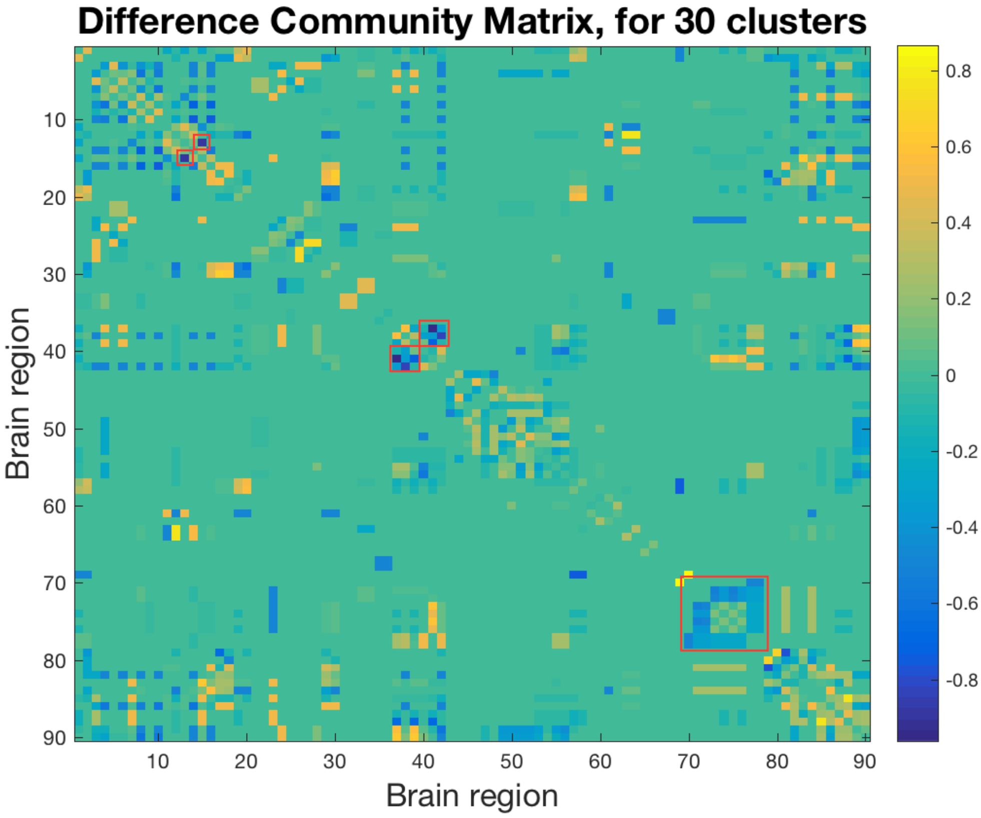


Figure 6: Difference community matrix for 30 clusters

brain regions belonging to the same cluster with higher K. To explain further, if we assigned K to the number of brain regions, every region would belong to its own cluster and we would only see a yellow diagonal and blue off diagonal elements. On the other hand, if we assigned K to 1, every region would belong to the same cluster. Therefore, we can imagine there is an ideal balance for the value of K which highlights only the most similar (connected) brain regions. So, as we increase from 10 to 30 clusters, we filter out regions with weak connections and only the strong ones remain. This is useful for determining a pattern across the healthy and diseased patients as there is more distinction between the two classes.

Additionally, we can see a common pattern of connected brain regions as the number of clusters increase, albeit, it becomes more defined when K is 30.

6 Comparison of community matrices

This section is a comparison the community matrices for the healthy controls and the diseased patients, namely, the ones shown in 5.

When looking at Figure 5a, the healthy controls show a stronger coherence in neuro-activities across the two hemispheres surrounding the diagonal as depicted by the bright yellow squares. Interestingly in Figure 5b, whilst the coherence is weaker, there are more connections in total as shown by the darker yellows highlighted in the red boxes. These could indicate that the disease weakens the connections between brain regions but allows more regions to connect. This may explain the adverse effects of the disease, as regions are functioning together when they never were originally.

7 Further Comments

In this section we look at finding the optimum number of clusters for identifying strong connectivity between brain regions.

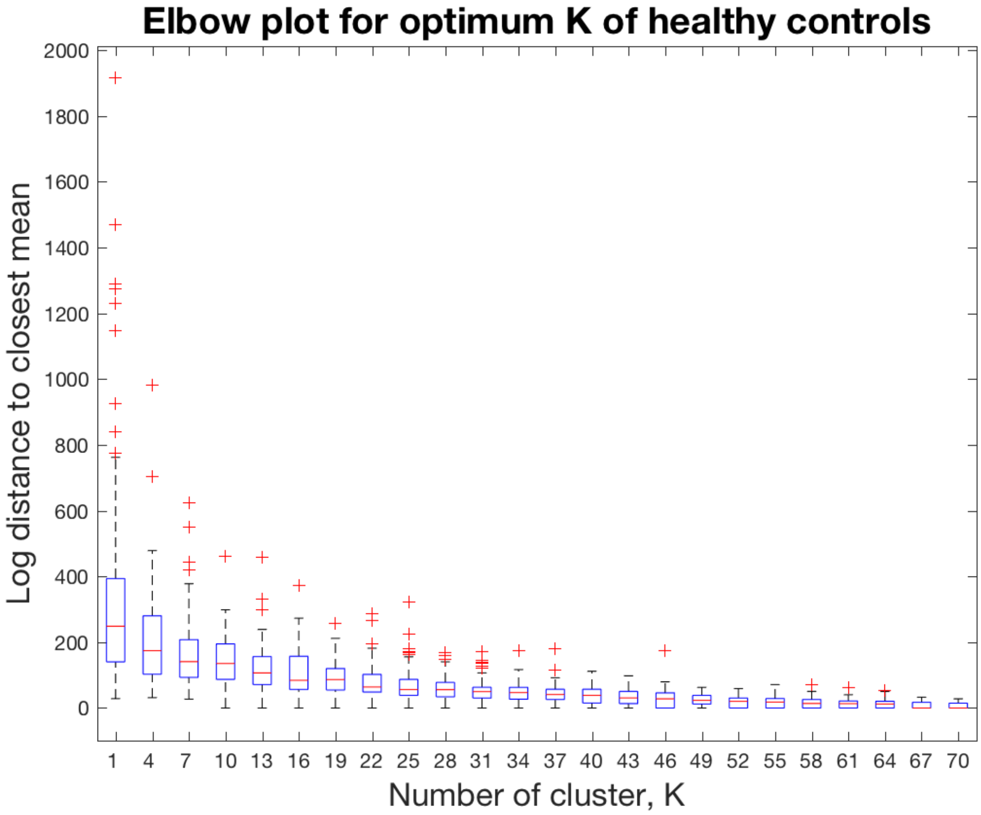


Figure 7: Elbow plot for healthy controls

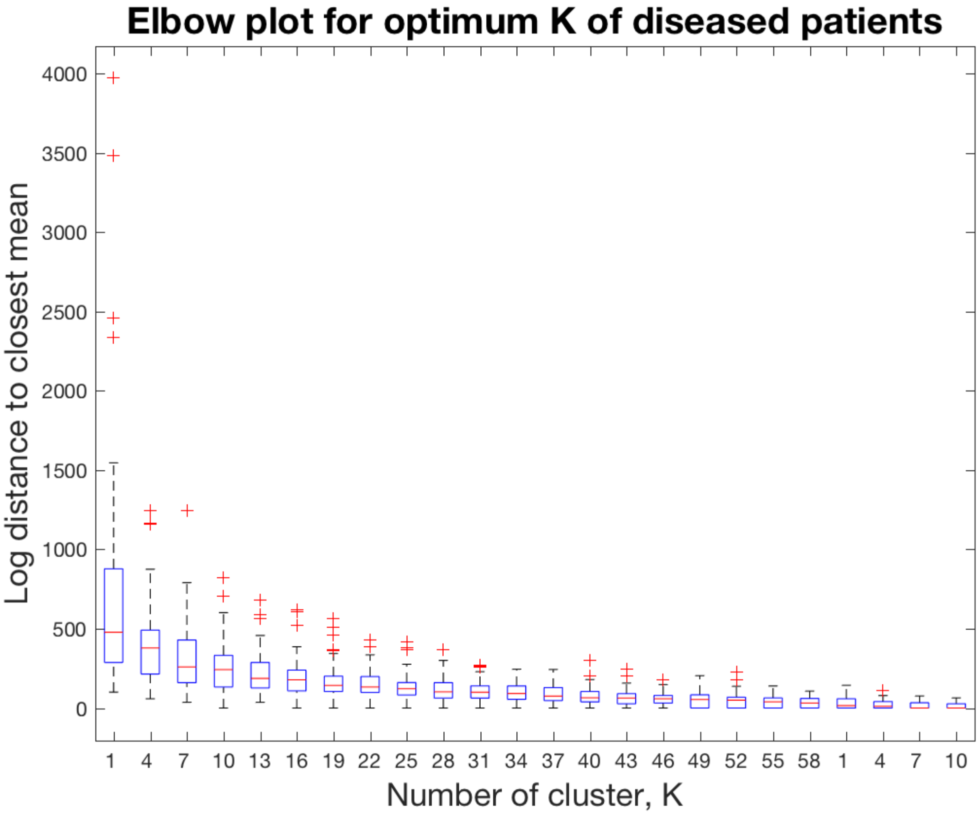


Figure 8: Elbow plot for diseased controls

8 References

[1] Jie Zhang, Wei Cheng, ZhengGe Wang, ZhiQiang Zhang, WenLian Lu, GuangMing Lu, and Jianfeng Feng. "Pattern Classification of Large-Scale Functional Brain Networks: Identification of Informative Neuroimaging Markers for Epilepsy". In: *PLOS ONE* 7.5 (May 2012), pp. 1–11. doi: 10.1371/journal.pone.0036733. url: http://dx.doi.org/10.1371%5C%2Fjournal.pone.0036733.

[2] Simon Rogers and Mark Girolami. *A First Course in Machine Learning*. 1st. Chapman & Hall/CRC, 2011. isbn: 1439824142, 9781439824146.

Appendix A: k-means clustering

```
%% Preamble
clear all; close all; clc;
savePlots = 0;
pool = parpool;
stream = RandStream('mlfg6331_64');
options = statset('UseParallel',1,'UseSubstreams',1,'Streams',stream);

%% Load the data
load('cbt3data.mat');
X = diseased(:,:,1)';
Y = healthy(:,:,1)';

%% Initialise Km
Km_diseased = zeros(size(X, 1));
Km_healthy = zeros(size(Y, 1));
person_type = {'DiseasedPatients';'HealthyControls'};

%% We get a community matrix for 10,20, and 30 clusters
for K = 10:10:30;
    %% Get average community matrix for all (20) patients
    tic;
    for person = 1:size(diseased,3); % iterate over each person and take average

        % we get the ith diseased and healthy person
        X = diseased(:,:,person)';
        Y = healthy(:,:,person)'; % we get the ith healthy person

        % For the ith person, we get the cluster index for each brain region
        [ClusterIndex_diseased] = kmeans(X,K, 'Replicates',500, 'Options', options)
    ;
        [ClusterIndex_healthy] = kmeans(Y,K, 'Replicates',500, 'Options', options);

        % We compare the cluster index of the ith brain region with every
        % other brain region index, we return 1 if identical, 0 if not.
        % This produces a 90 by 90 matrix (90 brain regions)
        for i = 1:90;
            Km_diseased(:,i) = double(repmat(ClusterIndex_diseased(i,1),90,1) ==
    ClusterIndex_diseased);
            Km_healthy(:,i) = double(repmat(ClusterIndex_healthy(i,1),90,1) ==
    ClusterIndex_healthy);
        end

        % We initialise the mean community matrix
        if (person == 1);
            meanKm_diseased = Km_diseased;
            meanKm_healthy = Km_healthy;
        end

        % We calculate the running mean of the diseased community matrix
        Km_3d_diseased = cat(3,meanKm_diseased, Km_diseased);
        meanKm_diseased = mean(Km_3d_diseased,3);

        % We calculate the running mean of the healthy community matrix
        Km_3d_healthy = cat(3,meanKm_healthy, Km_healthy);
        meanKm_healthy = mean(Km_3d_healthy,3);
    end
```

```
    %% View matrix as image (community matrix)
    meanKm_diseased_healthy = cat(3, meanKm_diseased, meanKm_healthy);

    for personType = 1:2
        figure(personType);
        imagesc(meanKm_diseased_healthy(:,:,personType));
        ti = sprintf('%s Community Matrix, for %g clusters',person_type{personType
    }, K);
        xlabel('Brain region','fontsize',16);
        ylabel('Brain region','fontsize',16);
        title(ti, 'fontsize',18);
        colorbar('eastoutside');

        if (savePlots == 1)
            name = sprintf('%s%g',person_type{personType},K);
            filename = sprintf(strcat(name,'.png'));
            saveas(gcf,filename);
        end

        figure(3)
        imagesc(meanKm_healthy - meanKm_diseased);
        ti = sprintf('Difference Community Matrix, for %g clusters', K);
        xlabel('Brain region','fontsize',16);
        ylabel('Brain region','fontsize',16);
        title(ti, 'fontsize',18);
        colorbar('eastoutside');

        if (savePlots == 1)
            name = sprintf('diffCommunityMatrix%g',K);
            filename = sprintf(strcat(name,'.png'));
            saveas(gcf,filename);
        end

    end
    toc;
end
```

```
%% Stops pool of workers
delete(gcp)
```

Appendix A: Elbow plot for optimising the number of clusters

```
%% Preamble
clear all; close all; clc;
savePlots = 0;
```

```
%% Load the data
load('cbt3data.mat');
```

```
%% Initialise Km and meanKm
person_type = {'Diseased Patients';'Healthy Controls'};
D_diseased_total = [];
D_healthy_total = [];
```

```
%% We get a community matrix for 10,20, and 30 clusters
for K = 1:3:70;
    tic;
    % we get the ith diseased and healthy person
    X = diseased(:,:,1)';
    Y = healthy(:,:,1)'; % we get the ith healthy person

    % For the ith person, we get the cluster index for each brain region
    [~,~,~,D_diseased] = kmeans(X,K, 'Replicates',50);
    [~,~,~,D_healthy] = kmeans(Y,K, 'Replicates',50);

    % We take the distance closest to a cluster
    D_diseased_min = min(D_diseased,[],2);
    D_healthy_min = min(D_healthy,[],2);

    % We store the ditances for each point to its respective mean
    D_diseased_total = [D_diseased_total, D_diseased_min];
    D_healthy_total = [D_healthy_total, D_healthy_min];

end
```

```
%% Boxplot of data
figure(1)
boxplot(D_diseased_total);
labels = (1:3:60);
set(gca, 'XTickLabel', labels);
% set(gca,'YScale','log')
xlabel('Number of cluster, K','fontsize',16);
ylabel('Log distance to closest mean','fontsize',16);
title('Elbow plot for optimum K of diseased patients', 'fontsize',18);
if (savePlots == 1)
    filename = ('elbowDiseased.png');
    saveas(gcf,filename);
end
```

```
figure(2)
boxplot(D_healthy_total);
labels = (1:3:90);
set(gca, 'XTickLabel', labels);
set(gca, 'XTickLabel', labels);
% set(gca,'YScale','log')
xlabel('Number of cluster, K','fontsize',16);
ylabel('Log distance to closest mean','fontsize',16);
title('Elbow plot for optimum K of healthy controls', 'fontsize',18);
if (savePlots == 1)
    filename = ('elbowHealthy.png');
    saveas(gcf,filename);
end
```