Pattern Recognition Practical 6

Group 24: Maikel Withagen (s1867733) Steven Bosch (s1861948)

October 21, 2015

Assignment 1

1

When we take $k = 2$ the Minkowski metric is the same as the Euclidean distance between the points, which is used as error function in other clustering methods such as K-means clustering.

2

See section A in the appendix for our implementation.

3

Looking at figure 1a we can clearly distinguish four main clusters within the data. The figures show that for higher values of t more connections are plotted. This is logical since a higher threshold permits higher distances between two points to be plotted, which overall results in more plotted connections. As for the optimal value of t, 0.05 is clearly too low, because we can see some connections within the clusters, but a lot of points that clearly belong to the clusters are left out because their distance to the other points is too high (see figure 1b). We can still see this for a t of 0.1, but on a smaller scale (see figure 1c). On the other hand a t of 0.25 is clearly too high, because multiple different clusters get connected through outliers, causing multiple clusters to be clustered together (see figure 1f). The same thing happens with a t of 0.2, where the two left clusters are connected (see figure 1e). Since this does not happen for a t of 0.15 and almost all of the points are assigned to a cluster (see figure 1d), this seems to be the optimal value of t. There is one point that does not get assigned to a cluster, so we will have to accept this as an outlier that does not belong to any cluster.

Assignment 2

1

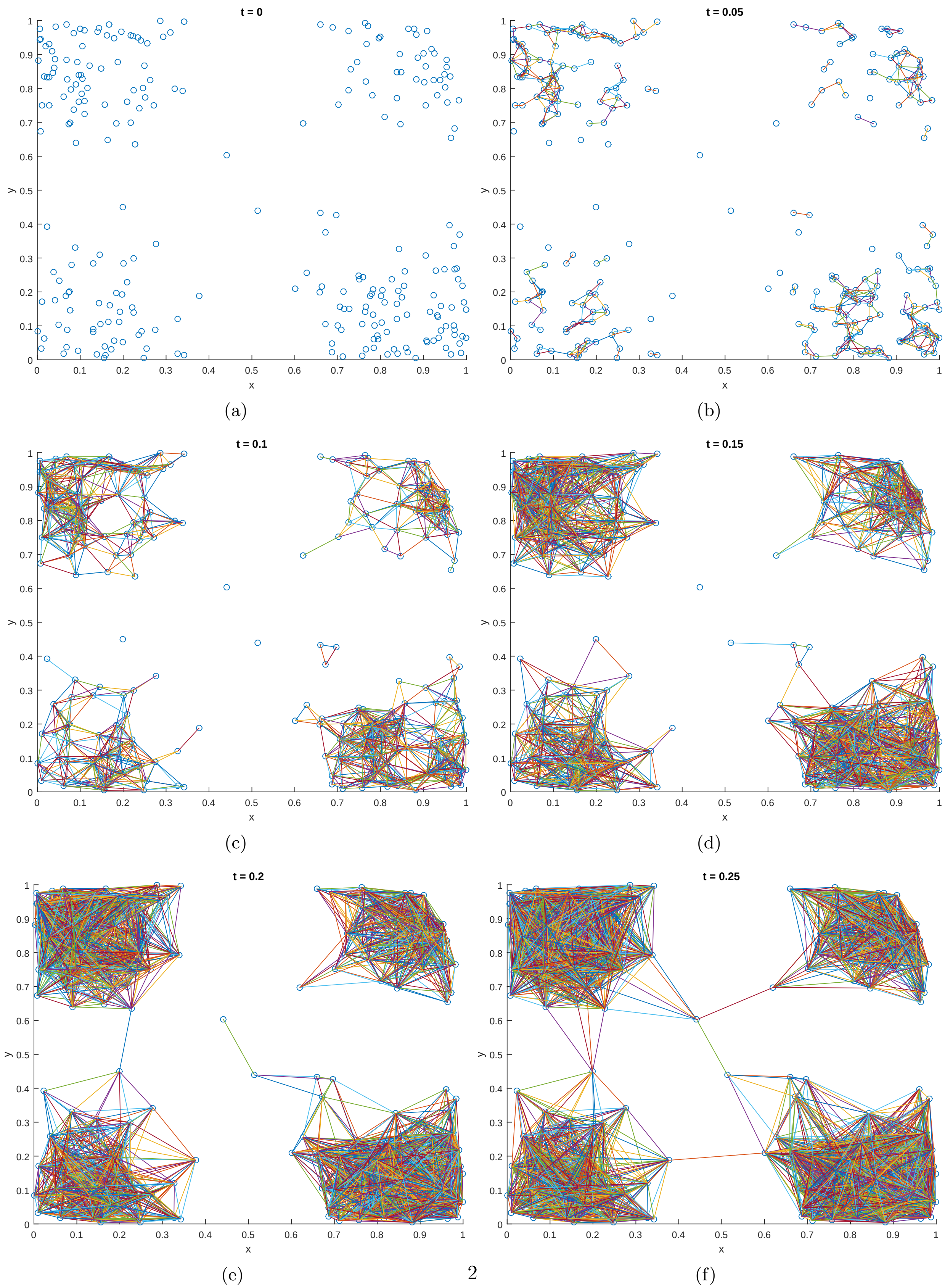
See section B in the appendix for our implementation.

2

Assignment 3

Using the code given in section C in the appendix, we computed the following J-values for the different clusterings:

Figure 1: Minkowski clustering for different threshold values.



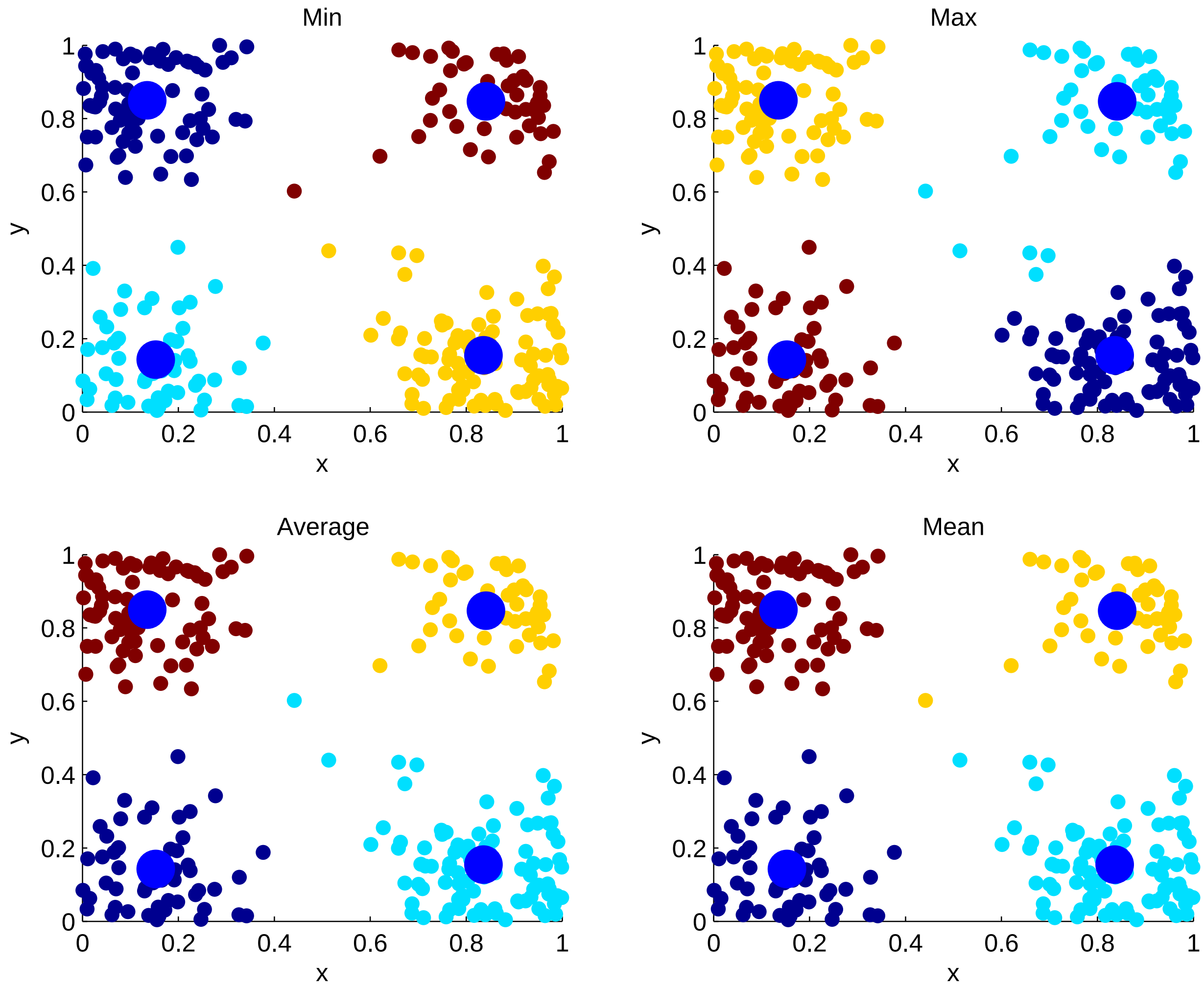


Figure 2: Agglomerative hierarchical clustering using different distance functions. The big circles are the centroids per cluster.

Table 1: J_e-values for different clusters

Clustering	J_e
$\{\{x1, x2, x3\}, \{x4, x5\}\}$	13.1667
$\{\{x2, x3, x5\}, \{x1, x4\}\}$	20.6667
$\{\{x4\}, \{x1, x2, x3, x5\}\}$	17.7500
$\{\{x4, x5\}, \{x1, x2, x3\}\}$	13.1667
$\{\{x3, x5\}, \{x1, x2, x4\}\}$	22.6667

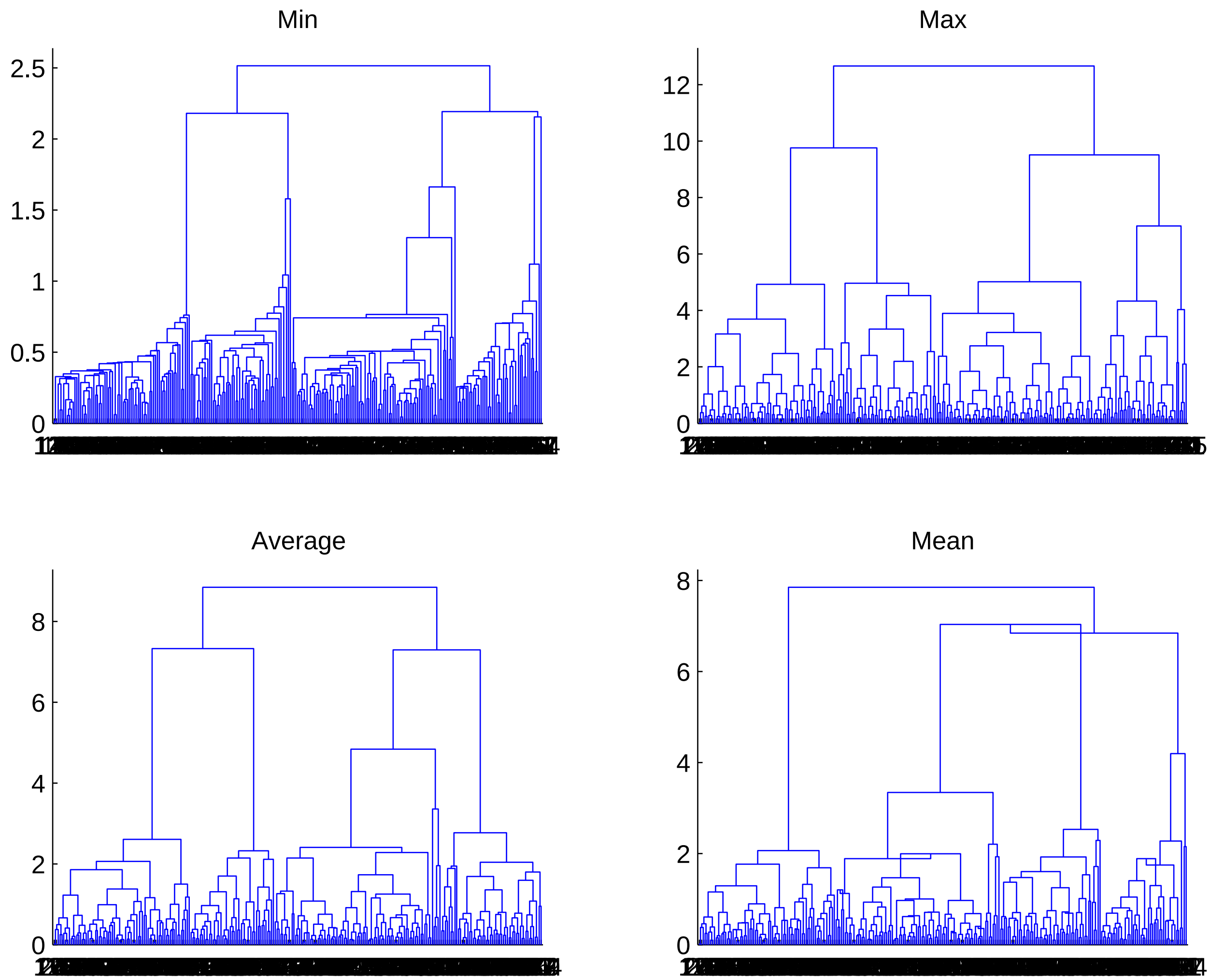


Figure 3: Dendroids for the agglomerative hierarchical clustering using different distance functions.

Appendix

A Assignment 1

../Code/Ass1.m

```
close all;
load('cluster_data.mat', 'cluster_data');
dat = cluster_data;
k = size(dat, 2);
% Calculate the minkowski distances between the points for k dimensions
dist = pdist2(dat, dat, 'minkowski', k);

% Make a new figure for every t-value and plot the relevant connections
for t = 0.00 : 0.05 : 0.25
    figure;
    hold on;
    plot(dat(:,1), dat(:,2), 'o');
    % Loop over all points and plot the connections when the distance
    % between two points is smaller than t
    for point = 1 : length(dat)
        for point2 = point+1 : length(dat)
            if dist(point, point2) < t
                plot([dat(point,1) dat(point2,1)], [dat(point,2) dat(point2,2)])
            end
        end
    end
    xlabel('x'); ylabel('y'); title(['t = ' num2str(t)]);
    print(sprintf(['../Report/Ass1_' num2str(t*100)]), '-depsc');
    hold off;
end
```

B Assignment 2

../Code/Ass2.m

```

X = cluster_data;
c_min = cluster(linkage(squareform(pdist(cluster_data)), 'single'), 'maxclust', 4);
c_max = cluster(linkage(squareform(pdist(cluster_data)), 'complete'), 'maxclust', 4);
c_avg = cluster(linkage(squareform(pdist(cluster_data)), 'average'), 'maxclust', 4);
c_mean = cluster(linkage(squareform(pdist(cluster_data)), 'centroid'), 'maxclust', 4);

hold on;
figure();
subplot(2,2,1)
scatter(X(:,1),X(:,2),[],c_min, 'filled');
hold on;
for group = 1:4
    plot(mean(X(c_min == group,1)),mean(X(c_min == group,2)), 'o', 'MarkerSize', 15, 'MarkerFacecolor', 'b');
end
xlabel('x');ylabel('y');title('Min');
subplot(2,2,2    )
scatter(X(:,1),X(:,2),[],c_max, 'filled')
hold on;
for group = 1:4
    plot(mean(X(c_min == group,1)),mean(X(c_min == group,2)), 'o', 'MarkerSize', 15, 'MarkerFacecolor', 'b');
end
```

```
xlabel('x');ylabel('y');title('Max');
subplot(2,2,3)
scatter(X(:,1),X(:,2),[],c_avg, 'filled')
hold on;
for group = 1:4
    plot(mean(X(c_min == group,1)),mean(X(c_min == group,2)), 'o', 'MarkerSize', 15, 'MarkerFacecolor', 'b');
end
xlabel('x');ylabel('y');title('Average');
subplot(2,2,4)
scatter(X(:,1),X(:,2),[],c_mean, 'filled')
hold on;
for group = 1:4
    plot(mean(X(c_min == group,1)),mean(X(c_min == group,2)), 'o', 'MarkerSize', 15, 'MarkerFacecolor', 'b');
end
xlabel('x');ylabel('y');title('Mean');
print(sprintf('../Report/Ass2_1'), '-depsc');

figure();
subplot(2,2,1); dendrogram(linkage(squareform(pdist(cluster_data)), 'single'), 270); title('Min');
subplot(2,2,2); dendrogram(linkage(squareform(pdist(cluster_data)), 'complete'), 270);title('Max');
subplot(2,2,3); dendrogram(linkage(squareform(pdist(cluster_data)), 'average'), 270);title('Average');
subplot(2,2,4); dendrogram(linkage(squareform(pdist(cluster_data)), 'centroid'), 270);title('Mean');
print(sprintf('../Report/Ass2_2'), '-depsc');
```

C Assignment 3

../Code/Ass3.m

```
x1 = [0 0];
x2 = [2 3];
x3 = [1 4];
x4 = [4 2];
x5 = [3 0];

% {{x1, x2, x3}, {x4, x5}}
m1 = 1/3 * (x1+x2+x3);
m2 = 1/2 * (x4+x5);
J1 = norm(x1-m1).^2+norm(x2-m1).^2+norm(x3-m1).^2+norm(x4-m2).^2+norm(x5-m2).^2

% {{x2, x3, x5}, {x1, x4}}
m1 = 1/3 * (x2+x3+x5);
m2 = 1/2 * (x1+x4);
J2 = norm(x2-m1).^2+norm(x3-m1).^2+norm(x5-m1).^2+norm(x1-m2).^2+norm(x4-m2).^2

% {{x4}, {x1, x2, x3, x5}}
m1 = x4;
m2 = 1/4 * (x1+x2+x3+x5);
J3 = norm(x4-m1).^2+norm(x1-m2).^2+norm(x2-m2).^2+norm(x3-m2).^2+norm(x5-m2).^2

% {{x4, x5}, {x1, x2, x3}}
m1 = 1/2 * (x4+x5);
m2 = 1/3 * (x1+x2+x3);
J4 = norm(x4-m1).^2+norm(x5-m1).^2+norm(x1-m2).^2+norm(x2-m2).^2+norm(x3-m2).^2

% {{x3, x5}, {x1, x2, x4}}
m1 = 1/2 * (x3+x5);
```

```
m2 = 1/3 * (x1+x2+x4);
J5 = norm(x3-m1).^2+norm(x5-m1).^2+norm(x1-m2).^2+norm(x2-m2).^2+norm(x4-m2).^2
```