Homework 2

1. Create problems and solutions on the course training wiki: http://gragg.math.kth.se/sf2524/merge_group_pages2.php?name=97359 corresponding to the current part of the course (block 2-3). This task is optional but can increase your bonus. Individual task (do work in groups). See how the work influences your bonus points under http://kth.instructure.com/courses/17791/pages/homework-slash-bonus-points-rules.

   Course training wiki: http://gragg.math.kth.se/sf2524/merge_group_pages2.php?name=63457

2. We want to carry out clustering on this data. Each row is one data point.

$$X = \begin{bmatrix} 1 & 1 & 1 \\ 1.5 & 2 & 0 \\ -0.5 & 0 & -3 \\ 0 & -0.5 & -1 \end{bmatrix}$$

   Carry out K-means on the data. (If you wish, you can use a computer to determine the distances (and the indicator vectors) but show the other computations by hand.)

   (a) Starting with R-vector

$$R = \begin{bmatrix} 3 & 3 & 3 \\ -1 & -1 & 0 \end{bmatrix}$$

   (b) Starting with R-vector

$$R = \begin{bmatrix} -0.5 & 0 & -3 \\ -1 & -1 & 1 \end{bmatrix}$$

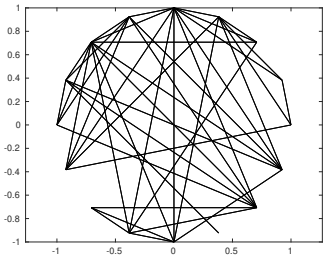
   Interpret similarities and differences between the solution in (a) and (b). Which solution is "best"?

Help similarity graphs in MATLAB: CANVAS → SF2526 → Pages → View all pages → Similarity graphs

3. We will now build a graph based on the following data:

```
n0=5; p=3;
randn('seed',0);
c=2;
A1=randn(n0,p)+c*[1,0,0];
A2=randn(n0,p)+c*[0,1,0];
A3=randn(n0,p)+c*[0,0,1];
A4=[0,0,0];
A=[A1;A2;A3;A4];
```

   Build a similarity graph using: ε-neighborhood for the standard Euclidean norm with $\varepsilon = 2.5$. Visualize it with circular node placement or matlabs node placement (as in the *help similarity graphs* in canvas). The circular placement should lead to a figure like the one to the right.



   (a) From the figure, determine how many connected components does the graph have.

   (b) Construct the unnormalized Laplacian. Run `eig(L)`. Relate the result in (a) with Lemma 2.3.3 in block2.pdf (or Proposition 2 in UvL).

   (c) Change the c-variable: $c = 3$. Provide a plot with the graph.

   (d) Repeat (a)-(b) for the new graph.

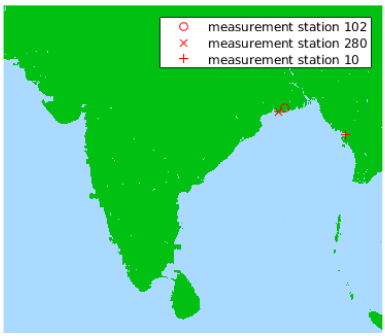
   (e) ** Another problem will be added here later **

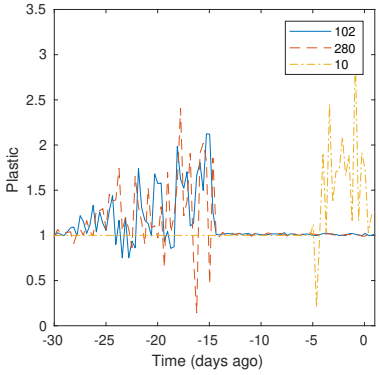
4. Do "Homework 2 quiz: computational spectral clustering" on CANVAS. This is mandatory and should be done individually (not in a group).

5. **Bengali cleanup simulation.** The Bengali bay countries want to carry out joint efforts against marine pollution, first by collecting information about plastic. They collect data in areas close to the shore for one month, in 937 locations. The goal is to establish seven regions where the plastic pollution is similar in order to coordinate efforts (such that the number of sampling points can be reduced and monitor the pollution easier over time).

   The time-series for three locations are visualized below. From the time-series data we see that the plastic pollution characteristics close to Kolkata are rather similar to each other, in comparison to the given location in Burma.

This exercise is inspired by marine pollution investigations on the Canary islands: *Microplastic and tar pollution on three Canary Islands beaches: An annual study*, Marine Pollution Bulletin, April 2018, Pages 494-502 https://www.sciencedirect.com/science/article/pii/S0025326X1730838X





The data is stored in `bengali_cleanup.mat` and the map over the area is stored in `bengali_map.png`. Variables in the mat-file:

- `x_coords` and `y_coords` map coordinates for the 937 locations.
- `timeseries`: A matrix where every row is the data from one measurement station, the corresponding time-points are given in `tv` (same for all stations).

The plot above can be obtained from:

```
load('bengali_cleanup.mat');
A=imread('bengali_map.png');
figure(1); clf;
imshow(A); hold on;
jv=[102,280,10];
plot(y_coords(jv),x_coords(jv),'r*') % Note: x and y reversed since images have swapped x and y axis.
figure(2);
plot(tv,timeseries(jv(1),:),'-'); hold on
plot(tv,timeseries(jv(2),:),'--')
plot(tv,timeseries(jv(3),:),'-.')
```

Only the time-series (not coordinates) will be used for the clustering. We will do sanity check with the coordinate vectors.

(a) Familiarize yourself with the data: Load the data and provide a map with all measurement stations.

(b) Construct a distance matrix from the time-series. Define the distance between two nodes (in the naive way) by computing the two-norm of the difference between the two time-series vectors. What is the time series distances between 102 and 280 and 10? Does it make sense?

(c) Construct a weight matrix by using kNN (version: or) with $k = 3$. Provide the graph (using MATLABs placement of nodes). Can a human identify the (seven) clusters?

(d) Carry out spectral clustering with seven clusters (RatioCUT version). You may use the matlab built-in function `kmeans` for the final step. Answer with seven maps, one for each cluster.

(e) Change the k in (c) to $k = 2$. Does the general conclusions change?

(f) (optional) Find hidden conclusions in the data. For instance, inland Sri Lanka seems to have similarities with some other part of Bengali bay, which one?