Math 285 HW4, Fall 2015.

Instructions:

- Due date: **Thursday, December 3, in class**.
- Please type your homework in Word or LaTex.
- All the data needed by this homework can be found on the course website: `http://www.math.sjsu.edu/~gchen/math285.html`.
- For each programming question, submit both the Matlab script and the output (any numerics and/or figures).

Questions:

(1) Consider a small toy data set *threelines.mat* that consists of 75 points along 3 lines in $\mathbb{R}^2$. This dataset was used in class to compute the full affinity matrix $\mathbf{A}_{\text{PC}} \in \mathbb{R}^{n \times n^2}$ ($n = 75$) based on the polar curvature. In this problem you are asked to try a different affinity measure that is based on the least squares fitting error

$$\mathbf{A}_{\text{LS}}(i,[j,k]) = \begin{cases} e^{-\frac{e^2_{\text{LS}}(\mathbf{x}_i,\mathbf{x}_j,\mathbf{x}_k)}{2\sigma^2}}, & \text{if all distinct} \\ 0, & \text{otherwise} \end{cases}$$

(a) Compute the matrix $\mathbf{A}_{\text{LS}}$ in full by MATLAB with $\sigma = .01$, and display it as an image. To reduce the computational load, use each pair (j,k) only once (i.e. use unordered pairs). The resulting $\mathbf{A}_{\text{LS}}$ is of size $n \times \binom{n}{2}$.

(b) Apply the Ncut algorithm with weights $\mathbf{W} = \mathbf{A}_{\text{LS}}\mathbf{A}_{\text{LS}}^T$ to find three clusters (display your result). What is the error rate?

(c) The mat file *threelines.mat* also contains the affinity matrix $\mathbf{A}_{\text{PC}}$ computed in class by using the polar curvature. Which affinity matrix is better and what is your criterion?

(d) Download the SCC package from `http://www.math.sjsu.edu/~gchen/scc.zip` and apply it to this dataset. What are the clusters it finds and how much error?

(e) Now add an outlier to the data set by modifying the first point:

$$X(1,:) = X(1,:) - [0, 0.5];$$

Recompute $\mathbf{A}_{\text{LS}}$ using your code above and implement the degree method in MATLAB to detect this outlier. Which points have the largest degrees?

(2) The Generalized PCA (GPCA) algorithm, available at `http://www.vision.jhu.edu/gpca.htm`, is an algebraic method for clustering linear subspaces, by playing with polynomials. This problem is meant to help you understand the main idea of GPCA.

First, let us assume a data set consisting of n points that lie exactly on two lines in $\mathbb{R}^2$, both passing through the origin (draw a picture to visualize the situation). We know that points on each line should satisfy an equation of the form $a_i x + b_i y = 0$ for some coefficients a_i, b_i, where $i = 1, 2$. So overall the points in the union of the two lines must all satisfy the following equation

$$(a_1 x + b_1 y)(a_2 x + b_2 y) = 0$$

which can be expanded to the form

$$c_{11}x^2 + c_{12}xy + c_{22}y^2 = 0.$$

GPCA starts by fitting such a quadractic polynomial to the given data to find the coeffecients c_{11}, c_{12}, c_{22}, and then factors this second-order polynomial into the two first-order equations in order to find the two lines. Of course, when the data is noisy, all the above equations will be only approximately true, but still one can find the best set of coefficients $\{c_{11}, c_{12}, c_{22}\}$ such that each of the following is as close to being an equality as possible[1]:

$$c_{11}x_i^2 + c_{12}x_iy_i + c_{22}y_i^2 \approx 0, \quad i = 1, \ldots, n.$$

Now, follow the above logic to answer the following questions:

(a) Suppose we need to cluster 3 lines in $\mathbb{R}^2$ (they all go through the origin), what kind of polynomial should we fit to the entire data set? Make sure you specify both the degree of the polynomial and also its terms (same below).

(b) Suppose we need to cluster 2 planes in $\mathbb{R}^3$ (they both go through the origin), what kind of polynomial should we fit to the given data?

(c) Suppose we need to cluster 2 nine-dimensional subspaces in $\mathbb{R}^{10}$ (they both go through the origin), what kind of polynomial should we fit to the given data?

(d) Suppose we need to cluster 2 arbitrary lines in $\mathbb{R}^2$ (they don't necessarily go through the origin), what kind of polynomial should we fit to the given data?

(e) Suppose we need to cluster 2 lines in $\mathbb{R}^{10}$ (they both go through the origin), how difficult is the problem?

[1]These n equations can be combined into one matrix equation

$$\begin{pmatrix} x_1^2 & x_1y_1 & y_1^2 \\ x_2^2 & x_2y_2 & y_2^2 \\ \cdots & \cdots & \cdots \\ x_n^2 & x_ny_n & y_n^2 \end{pmatrix} \begin{pmatrix} c_{11} \\ c_{12} \\ c_{22} \end{pmatrix} \approx \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

The best coefficient $(c_{11}, c_{12}, c_{22})^T$ is given by the smallest right singular vector of the matrix in front (why?).