Problem Set 13
(Spring '19)

Lectures 24, 25, 26

Because of the timing next week I will not ask you to turn in this problem set. But please be sure to go through these problems, as they are part of the syllabus of the course - suggestions will again be posted.

We have continued the discussion of dimensional reduction using principal component analysis into nonlinear analogs of principal components (projecting onto lower dimensional nonlinear surfaces). We have also discussed using graph structures on data sets (with graph vertices representing data points x_i edges representing pairs of 'nearby' data points) in order to take advantage of hidden lower dimensional structures. Spectral clustering is an example, in which the graph Laplacian matrix has eigenvectors which represent 'slowly varying' functions on the graph of data x_i.

A data point $X = (X_1, ..., X_p)$ has coordinates which are typically highly dependent on each other, and factor analysis tries to replace X by a function of it $S = (S_1, ..., S_p)$, with the S_i uncorrelated with each other. Independent component analysis tries to do better by making the S_i independent.

When the dimension p of the dataset greatly exceeds the data size N, new techinques collectively called *feature selection* or *dimensional reduction* become important. The idea is to reduce tens of thousands of features instead to (often) tens of features. Shrunken centroids is one of the techniques which has been popular in accomplishing this.

Problem 18.4 is related to the fact that even in high dimensional feature spaces it is possible to reduce a dataset with a small number N of elements down to an N dimensional feature subspace.

Reading: 14.7.1-14.7.3, 14.10,
18.1-18.3.

Problems: 14.14, 14.21, 18.2, 18.4 (Please look at/do these before taking the final).

Optional reading (not covered in class this year): 18.4-18.6 (not 18.6.1, 18.6.2), 18.7.1. We will not get to these topics in the course, but they are useful reading when you have the opportunity.