Problem set 1
Stat 154, SP 14
Due date: Friday, 2/14, 4pm

Problem 1
Use the stock market data. Work on the daily returns instead of the stock price - i.e instead of S_t, which is the time of a stock a time t, work on $R_t = S_t/S_{t-1} - 1$.

1. Do PCA on this data. Interpret the first few vectors of principal component loadings

2. Do hierarchical clustering of the stocks in the SP500, based on their daily returns. Comment briefly on the results. Explain briefly how you would use the results of the hierarchical clustering to display the covariance matrix of the stock returns. (You can use a subset of the data of course; just say which subset)

Problem 2
Use the 14 cancer microarray data at http://statweb.stanford.edu/ tibs/ElemStatLearn/

1. Perform PCA and kernel PCA (using a Gaussian kernel) on this data. Comment briefly on your findings.

2. Do K-means and K-medoids (using e.g PAM in R) on that data. Comment on the results.

Problem 3
Reproduce the simulation done in Figure 14.29 in the book. Do spectral clustering and kernel PCA on that data. Compare with usual PCA.

Problem 4
Create two simulations that allow you to visualize a good and a bad scree plot. Explain your reasoning and detail how you did the simulation. Plot the scree plots and comment.

Problem 5
Create simulations that allow you to investigate the effect of the ambient dimension on the stability and "faithfulness" of principal components. Describe your simulation setup in details and your conclusions.

Problem 6
Create a simulation setup in dimension $p = 20$ where the first four principal components individually explain roughly 20% of the variance, and the last 16 explain the remaining 20% of the variance. Repeat the experiment several times. Empirically, how stable is the first principal component? How stable is the subspace spanned by the first four. Explain what you've learned from that simulation.

Problem 7
Suppose that you deal with a large dataset drawn from a $\mathcal{N}(0, 1)$ distribution. What is IQR going to be? What about if you drew from a $\mathcal{N}(\mu, \sigma^2)$ distribution? Back to $\mathcal{N}(0, 1)$. What is the fraction of points who are going to be labeled outliers (to the right) with the standard boxplot method? What is the smallest number (depending on the sample size n) by which should you replace 1.5 in 1.5 IQR to guarantee that if the dataset consists of n points drawn independently from a $\mathcal{N}(\mu, \sigma^2)$ distribution, your whiskers will cover the minimum and the maximum of the sample (at least a.s as $n \to \infty$) with probability going to 1? How would your answer change if you wanted the expected number of points labelled as outliers to be less than c, where $c > 0$ is user defined.