STA 4102/5106: Homework Assignment #3

(Wednesday, September 10)
Due: Wednesday, September 17

1. **PCA and Linear Regression:** Consider a linear regression problem where $y \in \mathbf{R}^m$ and $X \in \mathbf{R}^{m \times n}$ are given, and we have to solve for the coefficients $b \in \mathbf{R}^n$ such that $||y - Xb||^2$ is minimized. In case n is too large to handle, we can use principal component analysis to reduce n to d and then solve for the coefficients.

(a) For the data provided on the website, first compute a matrix $X_1 \in \mathbf{R}^{m \times d}$ as follows: (i) Find the sample covariance matrix $C \in \mathbf{R}^{n \times n}$ of the elements of X, (ii) Compute the singular value decomposition (SVD) of C to obtain the orthogonal matrix $U \in \mathbf{R}^{n \times n}$, (iii) Set U1 to be the first d columns of U, and (iv) define $X_1 = XU_1 \in \mathbf{R}^{m \times d}$.

(b) Now solve for the coefficients $\hat{b}_1$ by minimizing $\| y - X_1 b_1 \|^2$.

(c) Compute the sum of squares of error, SSE = $\| y - X_1 \hat{b}_1 \|^2$.

For the dataset provided m = 200, n = 100, and use d = 10. Use "load hw3_1_data" command to load the data in Matlab to obtain X and y.

Compute and plot (use the command 'plot' in Matlab) the SSE for values of d ranging from 10 to 100 in the steps of 10. i.e. d = 10, 20, 30, …, 100.

2. PCA: In this problem we will find and display the principal direction of a 2D dataset.

(a) Take a matrix K = [1 0.8; 0.8 5]. Generate k = 100 observations of a bivariate normal random variable using:

x = sqrtm(K)*randn(2, k);

Plot this data on a 2D scatter plot using plot(x(1, :), x(2, :),'*').

(b) Perform PCA on this data with n = 2 and d = 1. Draw the dominant singular vector U1 on this plot using the command

plot(2*sqrt(S(1, 1))*[-U(1, 1) U(1, 1)], 2*sqrt(S(1, 1))*[-U(2, 1) U(2, 1)], 'g') ,

where [U, S, V] is the svd of the covariance matrix of the data.
Compute the variance in the first principal component. What is the ratio of this variance over the total variance in the original data?

3. PCA and Images: Consider the problem of analysis of images. Each (gray scale) image can be thought of as a matrix of numbers, say $I \in \mathbf{R}^{m_1 x m_2}$. We can rewrite this matrix as a long vector $X \in \mathbf{R}^{m_1 m_2}$. Setting $n = m_1 m_2$, we want to use PCA to reduce dimension from n to d. For the data file provided to you on the website perform PCA and present the following results:

(a) Show images of the first three principal directions of the data. That is, take the vectors U_1, U_2, and U_3 and display them as images. (Use the commands below to form images from vectors.)
(b) Take the first image in the data, and show its projection onto the principal subspace for d = 50 and d = 100. The projection of the first image into first d components is:

$$\sum_{i=1}^{d} (X(1,:) * U_i) * U_i.$$

Load the data file using "load hw3_3_data". This will give you a 200 × 644 matrix where each row of this matrix is a vector form of an image with $m_1 = 28$ and $m_2 = 23$. So there are 200 images in this dataset.
For a 644 length vector v you can form and display it as an image using:

```
I = reshape(v,28,23);
imagesc(I);
colormap(gray)
axis equal;
```

4. LDA (STA 5106 Students Only): Consider a labeled data set X with the following properties: there are m = 5 classes, each class has k = 10 observations, and each observation is a vector of size n = 3. Therefore, X can be thought of as three-dimensional array with dimensions 3 × 5 × 10. In matlab, X(:, i, j) denotes the jth observation vector of ith class.
Given this data, perform a linear discriminant analysis of the data for d = 1, and find the projection $U \in \mathbf{R}^{n \times d}$ that is optimal for separating observed classes. You can use the **eig** function in matlab to perform generalized eigen decomposition. For the resulting U:

(a) Plot the original data using command **plot3**.
(b) State U.
(c) Project the data X into Z, and plot the observations of Z.

Download X in "hw3_4_data" from the blackboard website.