STA 5106: Homework Assignment #3

(Thursday, September 12)
Due: Thursday, September 19

1. Write *multilinreg* program in Matlab to solve the following regression problem: find least-square estimate

$$\hat{b} = \arg\min_{b} \| y - Xb \|^2$$

where

$$X = \begin{bmatrix} 5 & 0 & 9 & 3 \\ 3 & 6 & 8 & 9 \\ 4 & 4 & 9 & 6 \\ 0 & 3 & 1 & 8 \\ 2 & 8 & 2 & 3 \end{bmatrix} \quad \text{and} \quad y = \begin{bmatrix} 20 \\ 17 \\ 32 \\ 10 \\ 12 \end{bmatrix}$$

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

(a) x is a matrix with size 2 by 18 in the following form:

$$\mathbf{x} = \begin{bmatrix} 15 & 16 & 12 & 14 & 13 & 15 & 16 & 21 & 12 & 11 & 19 & 14 & 13 & 14 & 16 & 17 & 12 & 16 \\ 13 & 11 & 13 & 12 & 9 & 14 & 12 & 16 & 9 & 8 & 15 & 13 & 15 & 13 & 12 & 16 & 11 & 9 \end{bmatrix}$$

The two rows are the experiment time costs (in minutes) of 18 students (first trial vs. second trial). Plot this data on a 2D scatter plot using plot(x(1, :), x(2, :),'*').

(b) Perform PCA on this data with sample size being 18 (i.e. each sample point is 2-dimensional). Draw the first principal direction on this plot. 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 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_3_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.

4, 5 (Optional): Use Python program to finish Problems 1, 2.

6 (Optional): Use Python program to finish Problems 3. In particular, use the following python commands to load the data in MAT format:

```
import scipy.io
mat_contents = scipy.io.loadmat('hw3_3_data.mat')
X = mat(mat_contents['X'])
y = mat(mat_contents['y'])
```

To use subroutines in mlr.py, you need to import this file with the following command:

```
from mlr import *
```