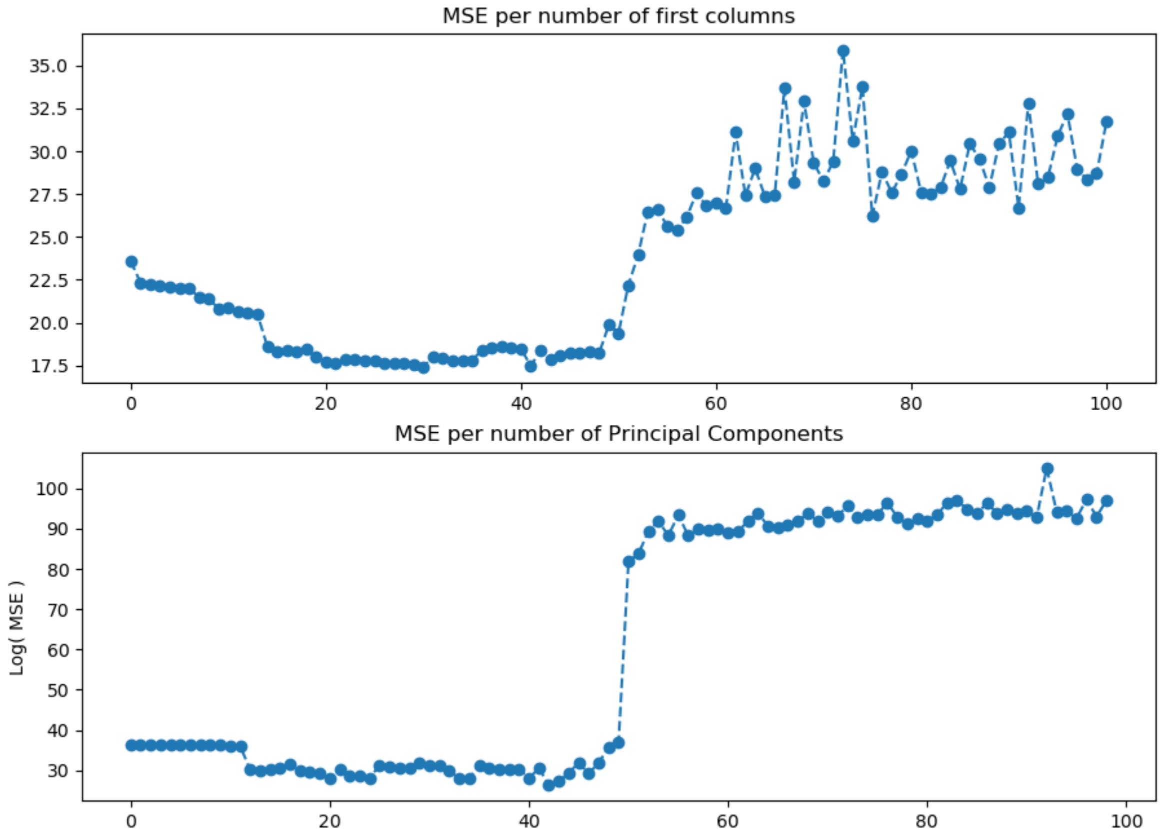
An Nguyen

EE 425 – Homework 4

Using PCA for Model Selection in Linear Regression

1. Simulated Data:

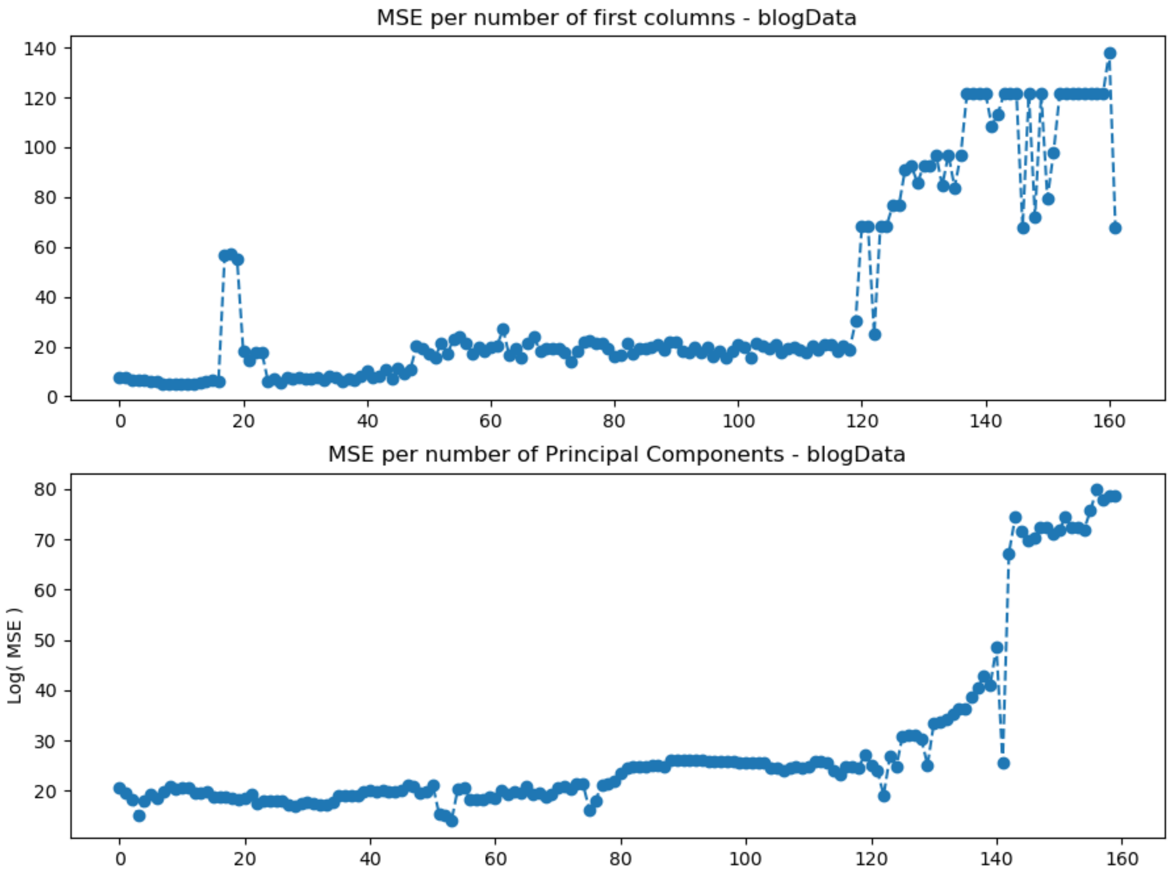


- The optimal number of features to be included is 31 first features
- The optimal number of principal components to be included is 43 components

From the graph we can see that as soon as n_row > n_col, the mean squared error skyrocketed (the y-axes are log – scaled), which is what we expected. From the start, as we add more and more features to the model, MSE decreases rapidly to a minimum before bouncing back. This is consistent with what we have seen in previous homework.

For PCA model, adding more components to the PCA space helps to represent the training data more closely, but also increase its bias.

2. blogData:



- The optimal number of features to be included is 11 first features
- The optimal number of principal components to be included is 54 components

To avoid singularity of the matrix (X’X), we removed all columns with identical values. The training data after cleaning: 248 observations of 163 features.

We perform the same procedure as in the simulation and got a similar result.