MATI 4211: Intelligent Systems Architectures– Summer 2017

Homework #1

Reports are due at Midnight on Monday, July 3rd
Please deposit your report in the digital drop box available on the course blackboard site.
Please remember to properly format for printing all documents submitted electronically.
Individual work will be strictly enforced!

Please install these R packages before proceeding:
zoo – time series analysis
plyr – data manipulation
MASS – stepwise regression
leaps – exhaustive search for subsets of features

Problem 1

In this problem, we apply Principal Component Analysis to stock market index data. A few useful functions for this sort of analysis are listed below.

```
prcomp(x, retx = TRUE, center = TRUE, scale = FALSE)
x: data frame; retx: return the projected x? ; center: center x beforehand? ;
scale: scale x? – if scale is TRUE, then PCA operates on the correlation matrix (instead of covariance)
obj <- prcomp(x)
obj$sdev: standard deviations of the sample PCs
obj$x: PC scores (in columns)
obj$rotation: PC weights (in columns), note: obj$x == x %*% obj$rotation
screeplot(obj, type = 'lines') # scree plot of data variance contained in the subsequent top PCs
```

We will use record of daily closing prices of S&P 500 stocks from January 1, 2011 to December 31, 2014 retrieved through Yahoo Finance. The data is stored in the attached file named “SP500_close_price.csv”. There are actually only 471 stocks in this file, the rest of the stocks were not included either because Yahoo Finance returned an error or because the stock was not listed as of January 1, 2011. The file named “SP500_ticker.csv” contains ticker information for each included stock, as well as the corresponding company name and its industry sector assignment. The attached R template file contains code to retrieve raw data. Feel free to try it out if you are curious. But you do not have to do so.

Items to address:

a) Fit a PCA model to log returns derived from stock price data. The code for deriving log returns is provided in the template file. Please use data frame named log.return.data as input to your PCA model. Having built the model, please do the following:
    1. Plot a scree plot which shows the distribution of variance contained in subsequent principal components sorted by their eigenvalues.

2. Create a second plot showing cumulative variance retained if top N components are kept after dimensionality reduction (i.e. the horizontal axis will show the number of components kept, the vertical axis will show the cumulative percentage of variance retained).
3. How many principal components must be retained in order to capture at least 80% of the total variance in data?
4. What is the magnitude of the estimated reconstruction error if we only retain top two of the PCA components?

b) Analysis of principal components and weights

1. Compute and plot the time series of the 1st principal component and observe temporal patterns. Identify the date with the lowest value for this component and conduct a quick research on the Internet to see if you can identify event(s) that might explain the observed behavior.
2. Extract the weights from PCA model for 1st and 2nd principal components.
3. Create a plot to show weights of the 1st principal component grouped by the industry sector (for example, you may draw a bar plot of mean weight per sector). Observe the distribution of weights (magnitudes, signs). Based on your observation, what kind of information do you think the 1st principal component might have captured?
4. Make a similar plot for the 2nd principal component. What kind of information do you think does this component reveal? (Hint: look at the signs and magnitudes.)
5. Suppose we wanted to construct a new stock index using one principal component to track the overall market tendencies. Which of the two components would you prefer to use for this purpose, the 1st or the 2nd? Why?