Homework Assignment #5

**Objective:** To gain experience in using alternative approaches to volatility calculation, neural network, and model comparison.

**Due Date:** Campus and weekend: May 21 and 22, respectively.

1. The file “d-gs9910.txt” contains daily open, high, low, closing prices and other variables for the stock of Goldman Sachs Group Inc. from May 4, 1999 to May 7, 2010. The names of the variables are given in the first row of the file (i.e. the header). The data are downloaded from Yahoo Finance. Use the data to construct the variance estimates $\sigma^2_{i,t}$ of Section 3.15.2 of the textbook for $i = 0, 1, 2, 3, 5,$ and $6$. Take the square root transformation to obtain volatility series. Multiple the volatility by $\sqrt{252}$ to obtain annualized volatility. Obtain mean, median, maximum, and minimum of each of the six annualized volatility series.

**Remark:** Yahoo puts the most recent observations first. To reverse the order, you may do the following in R.

```r
da=read.table('d-gs9910.txt', header=T)
T=nrow(da)
nc=ncol(da)
x=matrix(0,T,nc)
for (i in 1:T){
  ii=T-i+1
  x[i,]=da[ii,]
}
```

The matrix “x” contains the data in the usual time ordering. [The R command `rev` only works for a single series.]

2. Again, consider the data in “d-gs9910.txt”, but focus on the volatility of the log returns. Use the data and $n = 63$ to compute the Yang and Zhang (2000) variance estimate $\hat{\sigma}^2_{yz}$ of Section 3.15.2 of the textbook. Obtain a time plot of the estimated (and annualized) volatility series (square-root of annualized variance). Build a time series model to produce 1-step to 5-step ahead volatility forecasts at the forecast origin May 7, 2010. Hint: You can consider the log series of the volatility in modeling.
3. Consider the monthly U.S. unemployment rates from January 1948 to April 2010. There are 748 observations. Divide the data into estimation and forecasting subsamples with the former consisting of the first 650 observations. Use the out-of-sample root mean squared forecast errors to compare the following forecasting models:

(a) ARIMA\((2, 1, 1)(1, 0, 1)_{12}\), i.e. regular order \(c(2,1,1)\) and seasonal order \(c(1,0,1)\) with period 12.

(b) An AR\((11)\) model. (Selected by the command \texttt{ar}.)

(c) A neural network with \((x_{t-1}, x_{t-2}, x_{t-3}, x_{t-4}, x_{t-5}, x_{t-12}, x_{t-13}, x_{t-24})\) as input, a direct connection, and 2 nodes in the hidden layer, i.e. a 8-2-1 feedforward network.

Select the model that performed best in out-of-sample forecasts comparison to predict the US unemployment rate for May 2010 and June 2010 at the forecast origin April 2010.

4. Consider the tick-by-tick trade data of the Caterpillar stock from January 04 to January 08, 2010.

- Use the data within the normal trading hours only, i.e. from 9:30 am to 4:00 pm Eastern time, to construct a series of intraday 5-minute log returns. If there is no trading within a 5-minute interval, assume that the log return is zero. If there are multiple trades in a 5-minute interval, use the last trade to obtain the price for that interval. Plot the log return series.

- Are there any serial correlations in the intraday 5-minute log return series? Use \(Q(10)\) to perform the test.

[This problem requires some data processing in R.]

5. Again, consider the tick-by-tick trade data of the Caterpillar stock from January 4 to January 8, 2010.

- Construct the series of the number of trades within a 5-minute intervals. Use data in the normal trading hours only.

- Compute the ACF of the constructed time series, say from lag 1 to lag 250. Is there any evidence of diurnal pattern? [No formal test is needed. Simply comment on the ACF plot.]

**Reading assignments:** Chapters 3 and 4 of the textbook.