Assignment #2: Time-series Analysis and Pattern Finding using KNN
Spring 2016
CS3943/9223
Prof. Rumi Chunara
Total: 30 points + 2 possible bonus

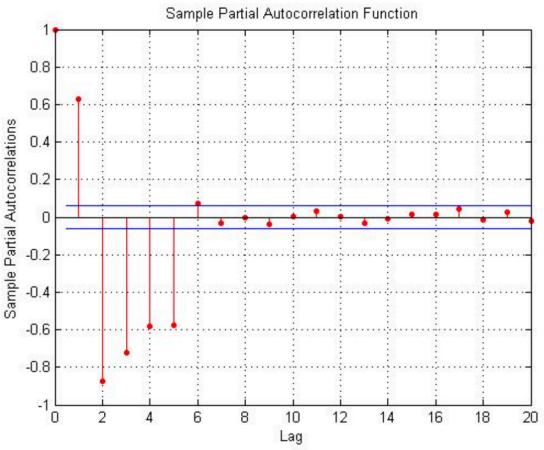
All questions must be completed in R. Implement and comment your code so that anyone reading the file can reproduce the code easily (e.g. set the file path once at the beginning of the script where it can be easily changed). Save the code as an R markdown file, and upload it to NYU classes.

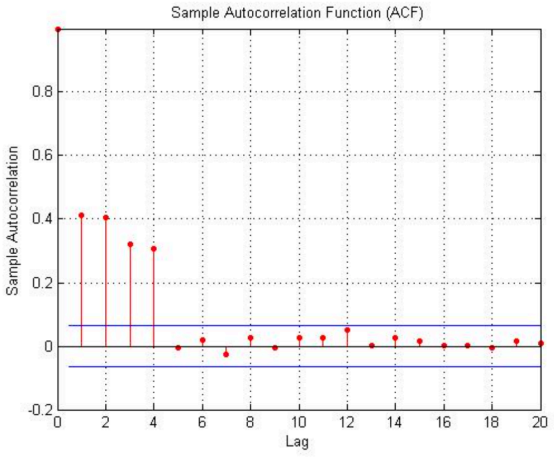
1. Time-series Analysis (15 points).

In class we discussed the Johnson & Johnson data set of quarterly earnings per share. This data set (JoJo.dat) is available on the NYU Classes site (in the Datasets folder) and should be used in this homework.

Forecasting based on ARIMA (autoregressive integrated moving averages) models, commonly know as the Box–Jenkins approach, comprises following stages: i.) Model identification ii.) Parameter estimation iii.) Diagnostic checking. We will accomplish those steps in this question.

a) An important step before fitting an ARIMA function is to make sure the timeseries is stationary. To do this, there are many ways to remove trend and seasonality. One easy approach is to use the function **ts( )** in the stats package to remove seasonality, and use diff to remove a trend. Use these functions (or other approaches if you wish) to make the JoJo series stationary. (3 points)
A first step in analyzing (stationary) time series is to examine the autocorrelations (ACF) and partial autocorrelations (PACF). R provides the functions **acf( )** and **pacf( )** for computing and plotting of ACF and PACF. The order of "pure" AR and MA processes can be identified from the ACF and PACF plots. Example ACF plots suggesting AR(5) and MA(4) values, left and right respectively, are shown below.





b) What order AR and MA are the JoJo.dat? Plot the ACF and PACF of the data. Indicate the units of lags in the plots. What kind of ARMA would you deem appropriate based on these plots? (3 points).
c) Once the order of the ARIMA(p,d,q)–model has been specified, the function **arima( )** from the stats package can be used to estimate the parameters: **arima(data,order=c(p,d,q))**. Try out different values of p and q (3 points)
d) A first step in diagnostic checking of fitted models is to analyze the residuals from the fit for any signs of non–randomness. R has the function **tsdiag( )**, which produces a diagnostic plot of a fitted time series model. Are the residuals stationary? (2 points).
e) Once a model has been identified and its parameters have been estimated, one purpose is to predict future values of a time series. Lets assume, that we are satisfied with the fit of an ARIMA(1,0,1)–model to the JoJo data using the **predict( )** function (2 points).
f) Plot the result with confidence intervals (2 points).

2. Pattern Finding (15 points total + 2 possible bonus).

Let's practice what we've learned for pattern finding using KNN and the Glass Identification dataset, which is available here with description: [http://archive.ics.uci.edu/ml/datasets/Glass+Identification].

Read the data into a DataFrame and briefly explore the data to make sure the DataFrame matches your expectations.

a) Now, let's perform a binary classification by transforming the data. Create a new DataFrame column called "bi". (2 points).
   - If the type of glass is one through four, set bi = 0.
   - If the type of glass is five through seven, set bi = 1.
b) Create a feature matrix "fea" using all features – this requires you to select carefully from all the data frame columns, and create a response vector "y" from the "bi" column (2 points).
c) Split *fea* and *y* into training and testing sets. How do you pick the size of each? (2 points).
d) Fit a KNN model on the training set using K=5. There are multiple functions in R available for this (2 points).
e) Make predictions for the testing set and calculate testing accuracy (2 points).
f) Write a loop that computes the testing accuracy for a **reasonable** range of K values (2 points).
g) Plot the K value versus testing accuracy to help you choose an optimal value for K. What is that optimal value? (3 points).
h) Calculate the testing accuracy that could be achieved by always predicting the most frequent class in the testing set. (This is known as the "null accuracy".) (2 points).

**Bonus: Explore the data to determine which features look like good predictors, and then redo this exercise using only those features to see if you can achieve a higher testing accuracy. (Up to 2 bonus points).