Analyzing SP 500 Index

using time series and machine learning methods

Yanni Ge Ying Shi Hairong Xie

University of California,Berkeley

March 17, 2014

Whose mind are we going to change? About what?

We are going to change/reinforce individual investors' mind about tracking and predicting SP500 index.

Data

SP500 data exacted from Yahoo Finance:

	AA	AAPL	ABC	ABT	ACE	ACN	ACT	ADBE	ADI
Date									
2010-01-05	15.38	206.05	24.95	22.90	43.80	38.89	39.89	37.70	27.96
2010-01-06	16.18	202.77	24.72	23.02	43.20	39.30	40.02	37.62	27.91
2010-01-07	15.84	202.40	24.32	23.21	43.45	39.27	39.70	36.89	27.69
2010-01-08	16.23	203.75	24.58	23.33	43.20	39.11	39.41	36.69	27.84
2010-01-11	16.64	201.95	24.86	23.45	43.67	39.07	39.77	36.21	27.69
2010-01-12	14.80	199.65	25.03	23.38	43.76	38.82	39.72	35.66	26.54
2010-01-13	15.24	202.47	25.52	23.61	43.90	39.27	40.89	36.28	26.53
2010-01-14	15.08	201.29	25.68	23.63	44.32	39.61	41.10	35.90	26.50
2010-01-15	14.91	197.93	25.40	23.69	43.85	39.33	40.90	35.87	25.61

LASSO

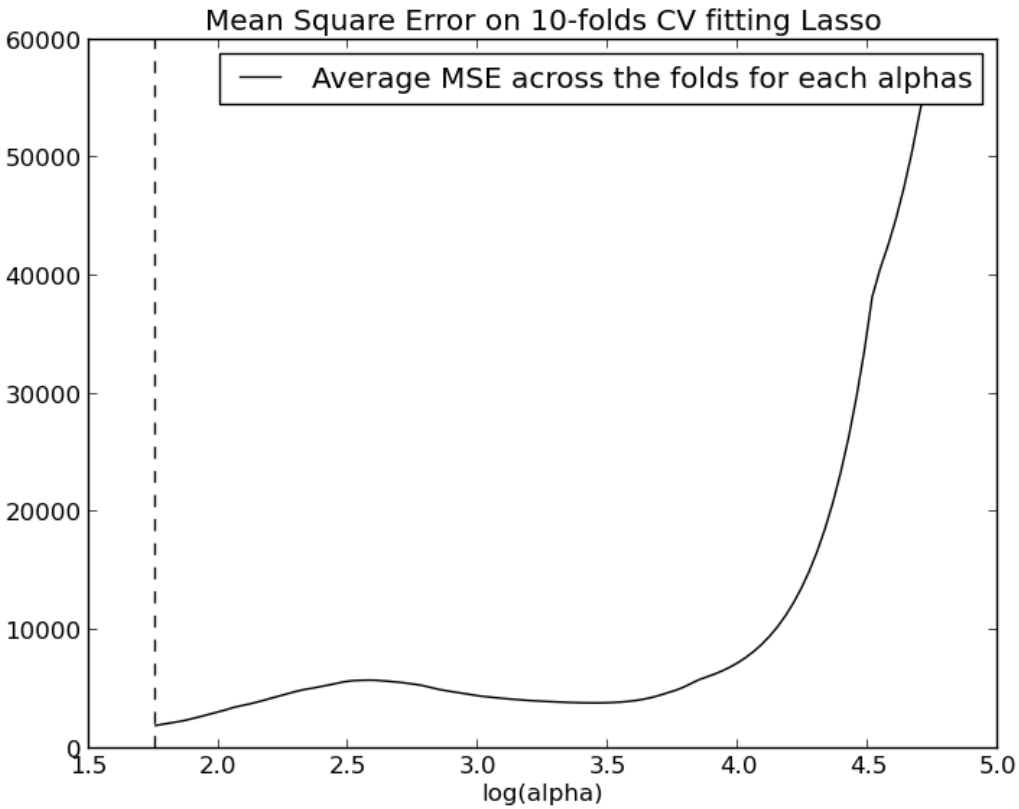

Figure 1: We determined the size of the penalty term alpha of LASSO by using 10-fold cross-validation. For each alpha, we calculated the average of the 10 MSE's across the folds and plotted it. We selected the alpha that gave us the least average MSE, and that LASSO picked 17 companies that will be plotted on the next page.

PCA

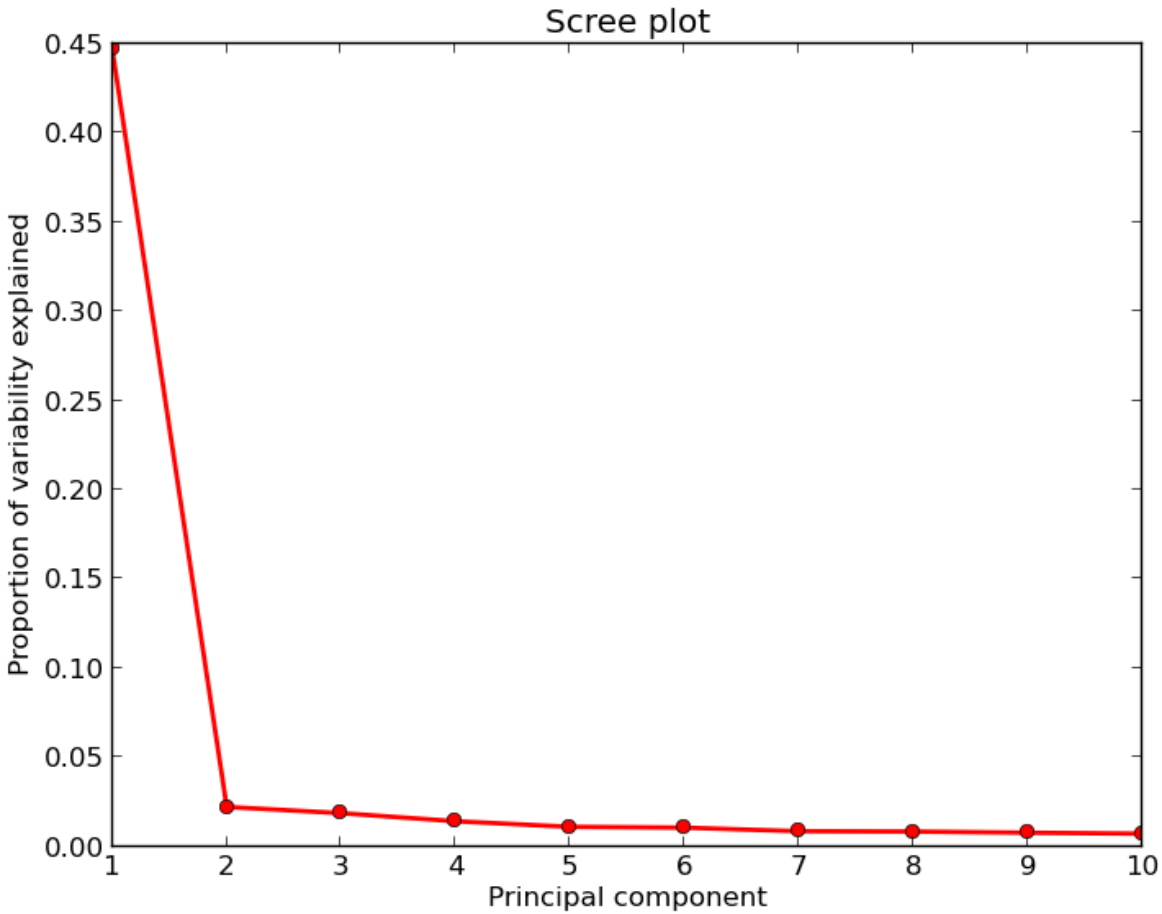

Figure 2: We applied PCA on daily returns. This scree plot shows the percentage of variability explained by each principal component. The first principal component explains around 45%.

PCA

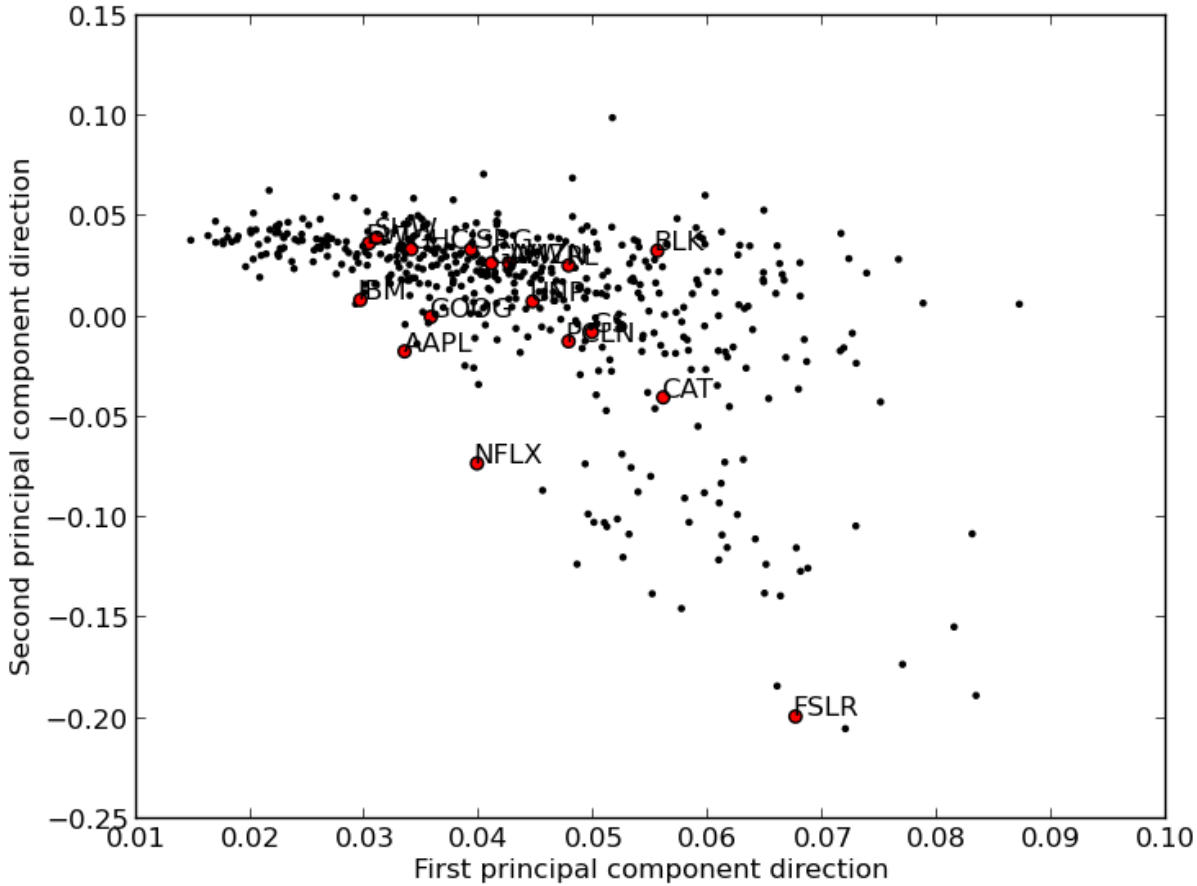

Figure 3: By plotting the first eigenvector against the second eigenvector, we can visualize the 446 companies with the selected 17 companies (red dots) labeled. The 17 companies seem to cover the trend of the whole.

PCA

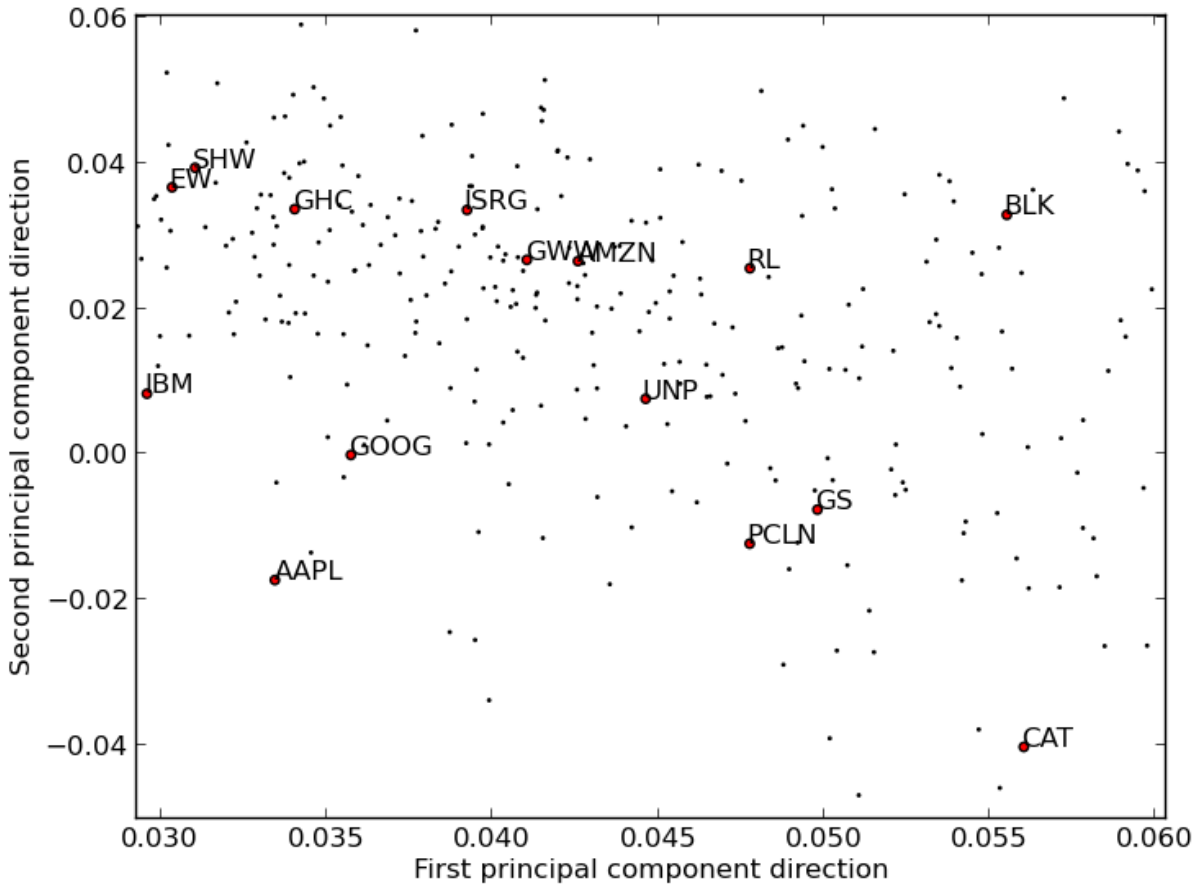

Figure 4: If we zoom in to the crowd, we can find clusters of stocks from different sectors, for example, APPLE, IBM and GOOGLE toward left side.

Density Plot

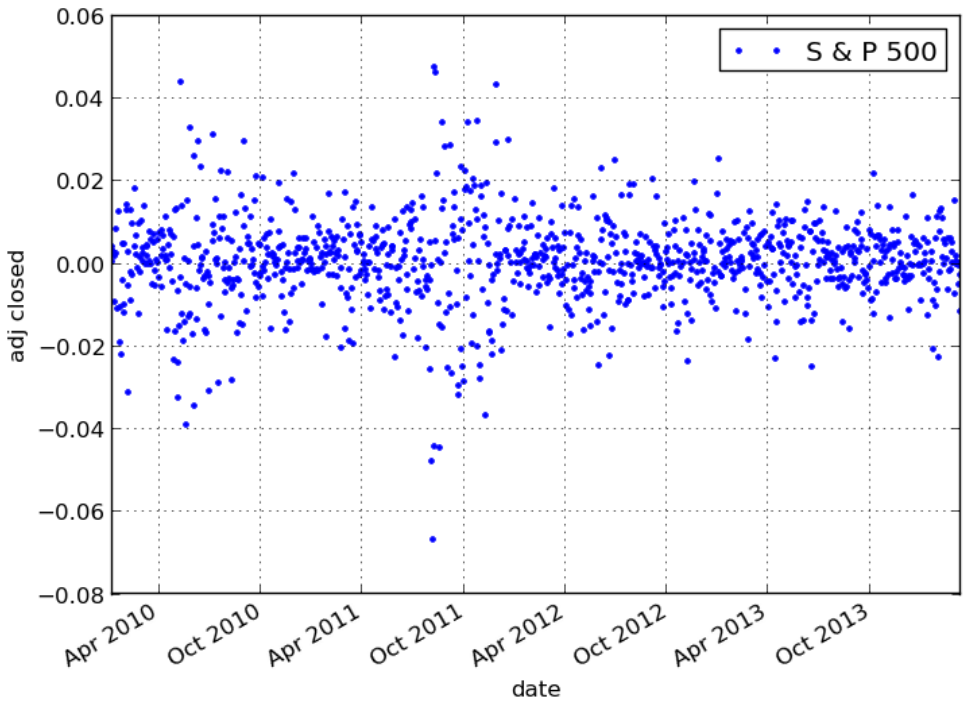

Figure 5: We calculated the daily return from adjusted closed price for SP 500 index. This is the scatter plot for these daily return from Jan 2010 to Mar 2014 (x-axis: time points; Y-axis: daily return). It seems that they are well distributed around 0 except that there are more variability during two time period: 1. May 2010 - Aug 2010; 2. Aug 2011 - Dec 2011. This inspires us to check what kind of distribution the daily return is.

Density Plot

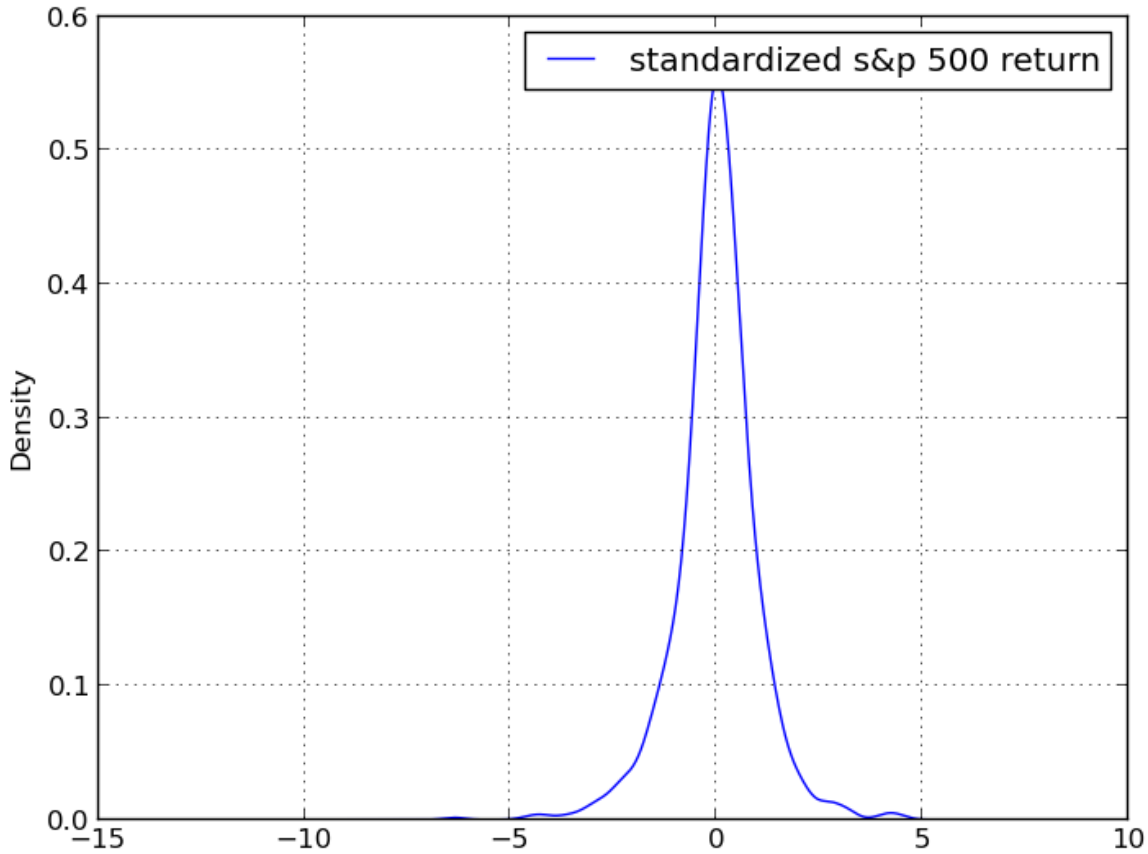

Figure 6: We plot the density of standardized daily return. The x-axis is the daily return after standardization; Y-axis is the density for the standardized daily return. The shape looks like normal distribution.

Density Plot

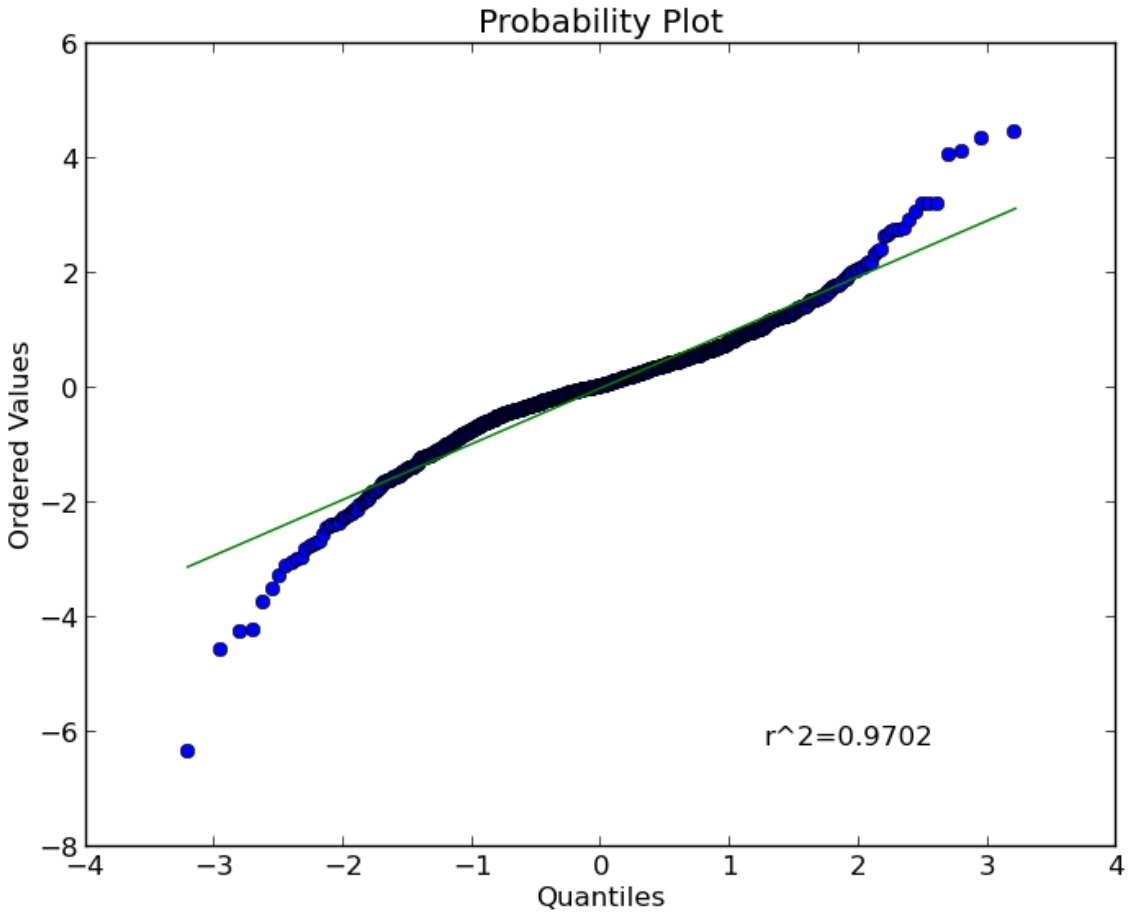

Figure 7: We plot the density of standardized daily return. The x-axis is the daily return after standardization; Y-axis is the density for the standardized daily return. The shape looks like normal distribution.

Density Plot

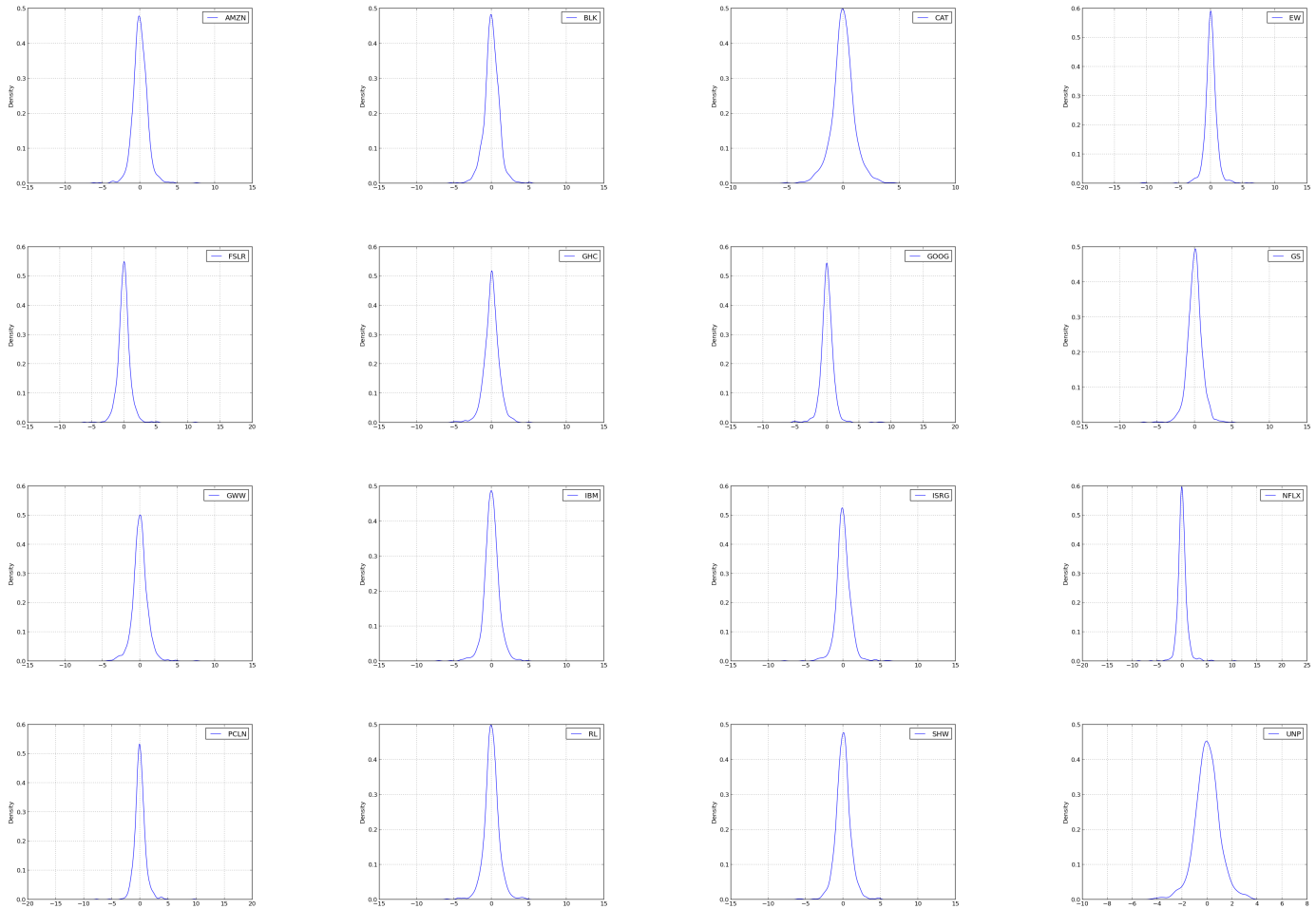

Figure 8: Remember we have 17 stocks picked out from PCA. We plot the density plots for the 16 stocks of them and the 17th stock - "AAPL" for APPLE is in the next page. They all seems to be bell shaped.

Density Plot

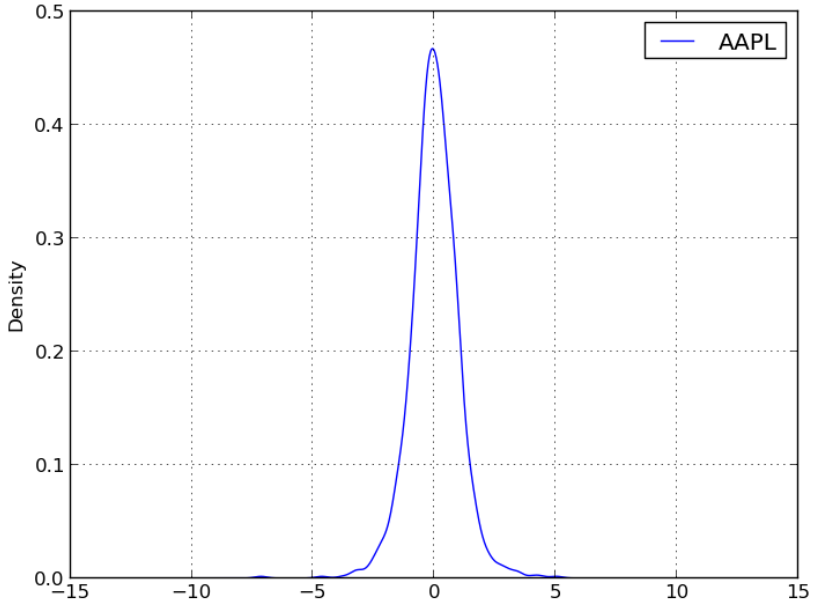

Figure 9: This is the density for APPLE stock. We use the normal distribution to make forecasting. Naturely, the mean daily return would be the best guess to make forecasting. For example, on Mar 14 the adjusted closed price is 525.69 and the mean daily return is 0.104%. Then the best guess for the adjusted closed price would be $524.69 * (1 + 0.104\%) = 525.24$. Best guess means this value would have largest probability to happen. And we can expect that the probability for the price to be larger than 525.24 is 0.5.

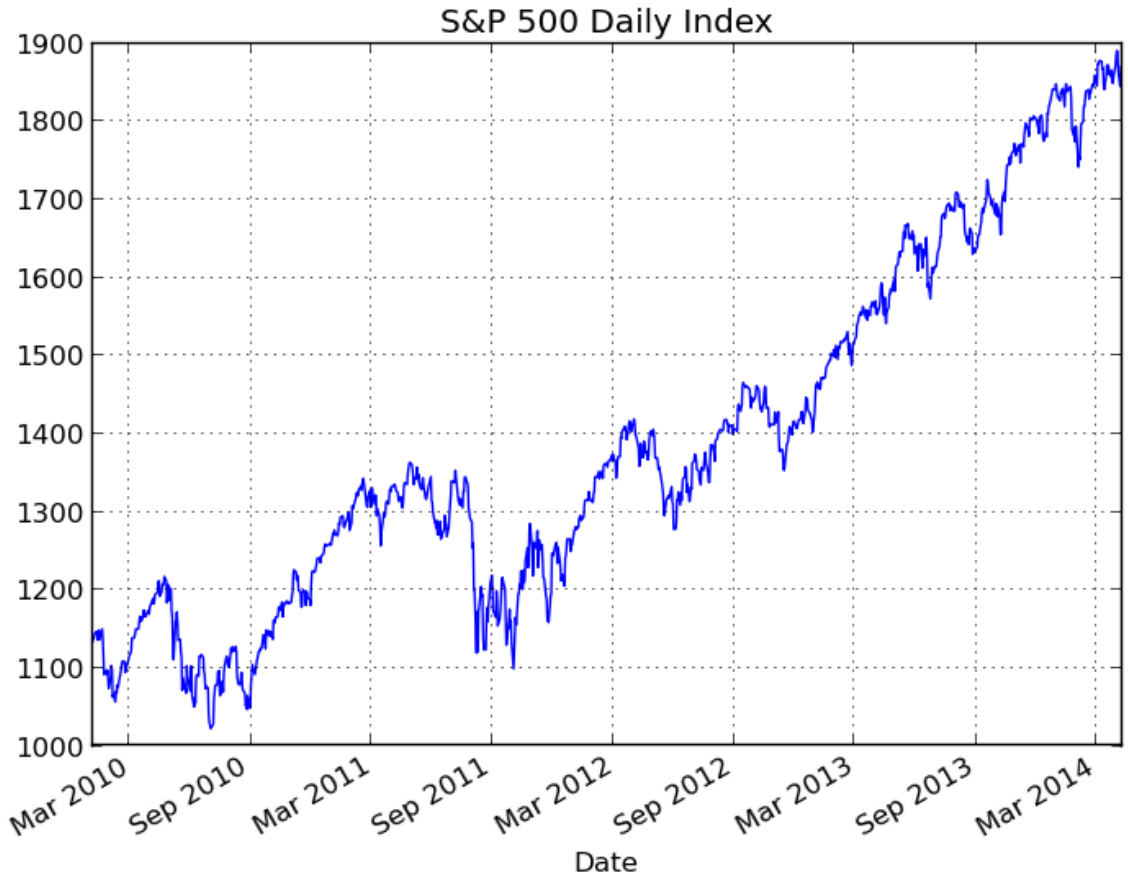

Figure 10: We continue with prediction by using time series method. We begin with plotting the SP500 daily prices from Jan 4, 2010 to now and have a basic idea of the trend.

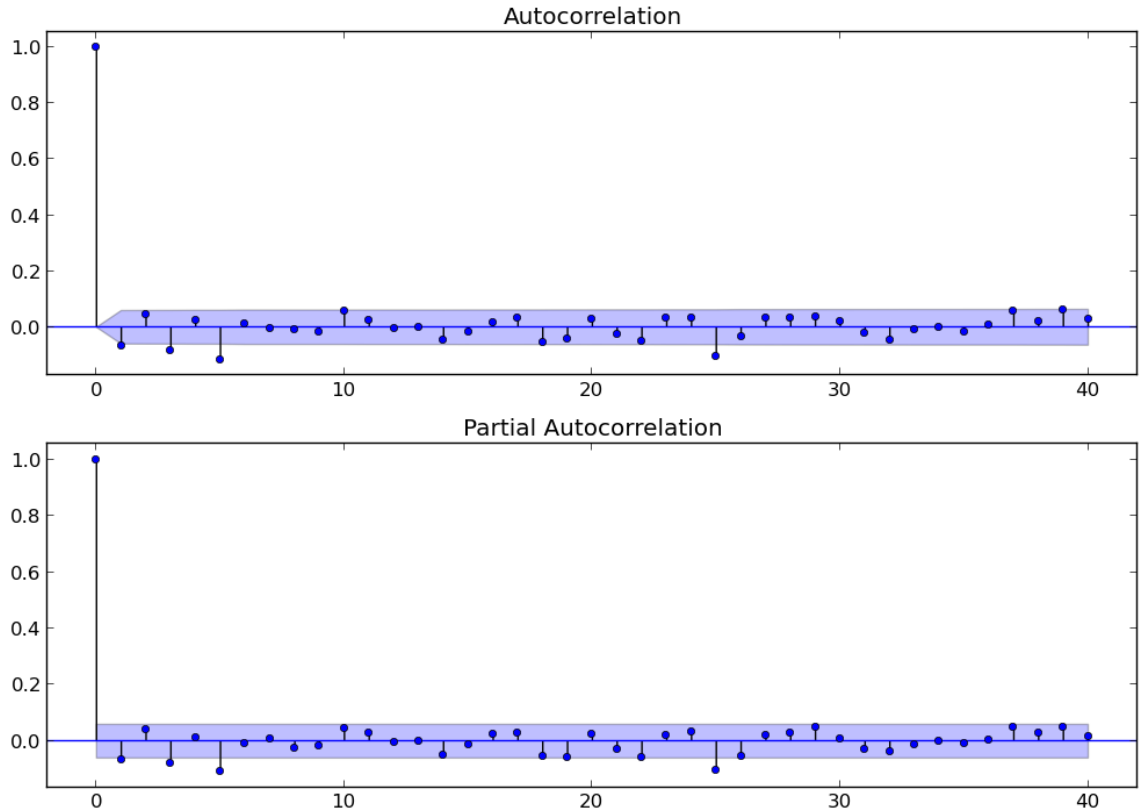

Figure 11: We applied Autocorrelation and Partial Autocorrelation function on SP500 daily excess returns, and tails are close to 0, which satisfied the stationary assumption.

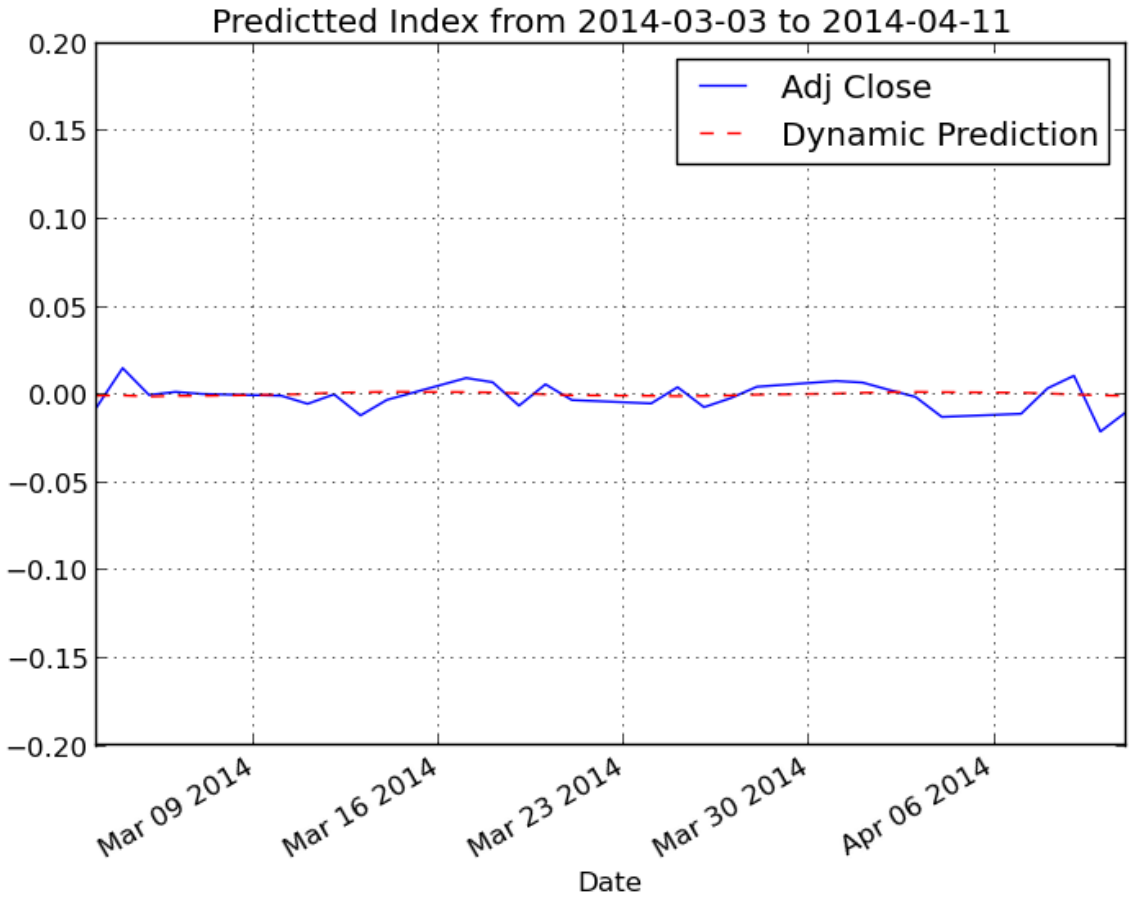

Figure 12: We applied ARIMA model (3,0,3) to predict more than 1 month daily returns. The blue line is the true value, and the red dashed line is the prediction, which are found to be pretty close to each other.

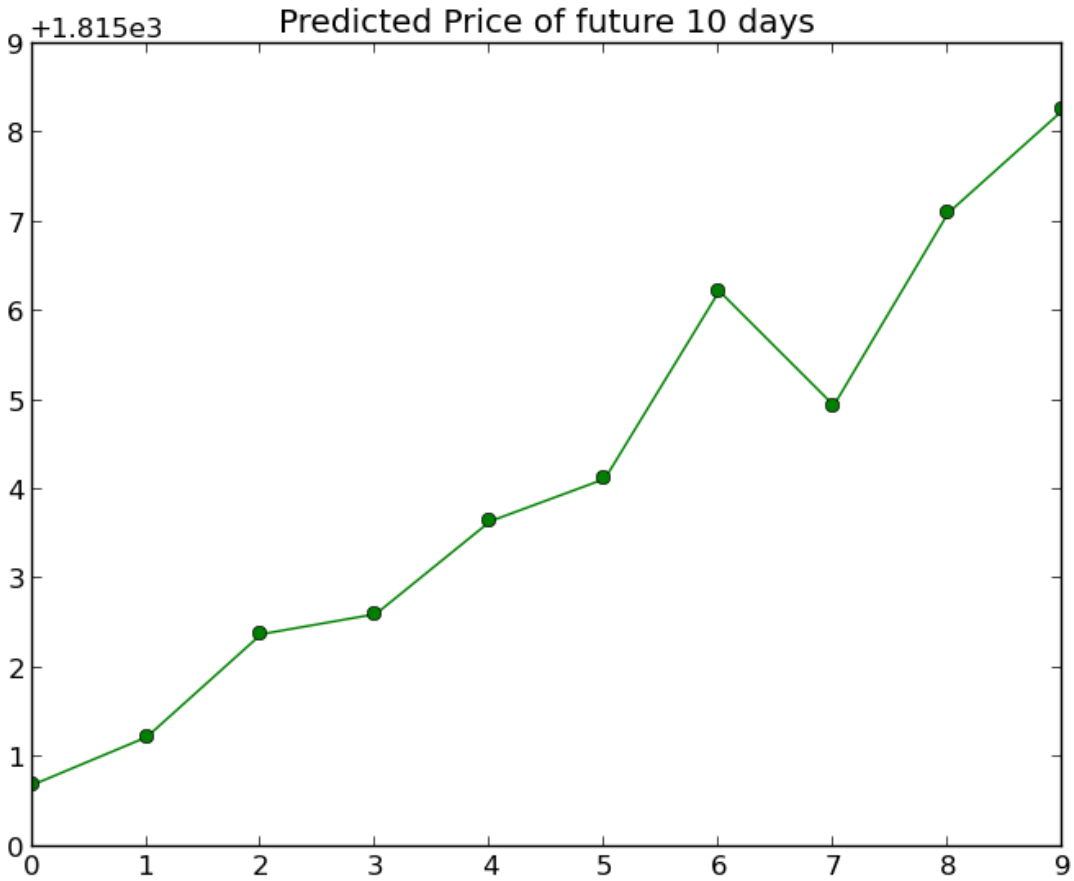

Figure 13: We make a transformation from daily returns to predicted prices of the next 10 days from 04/11/2014 to visualize the index trend.

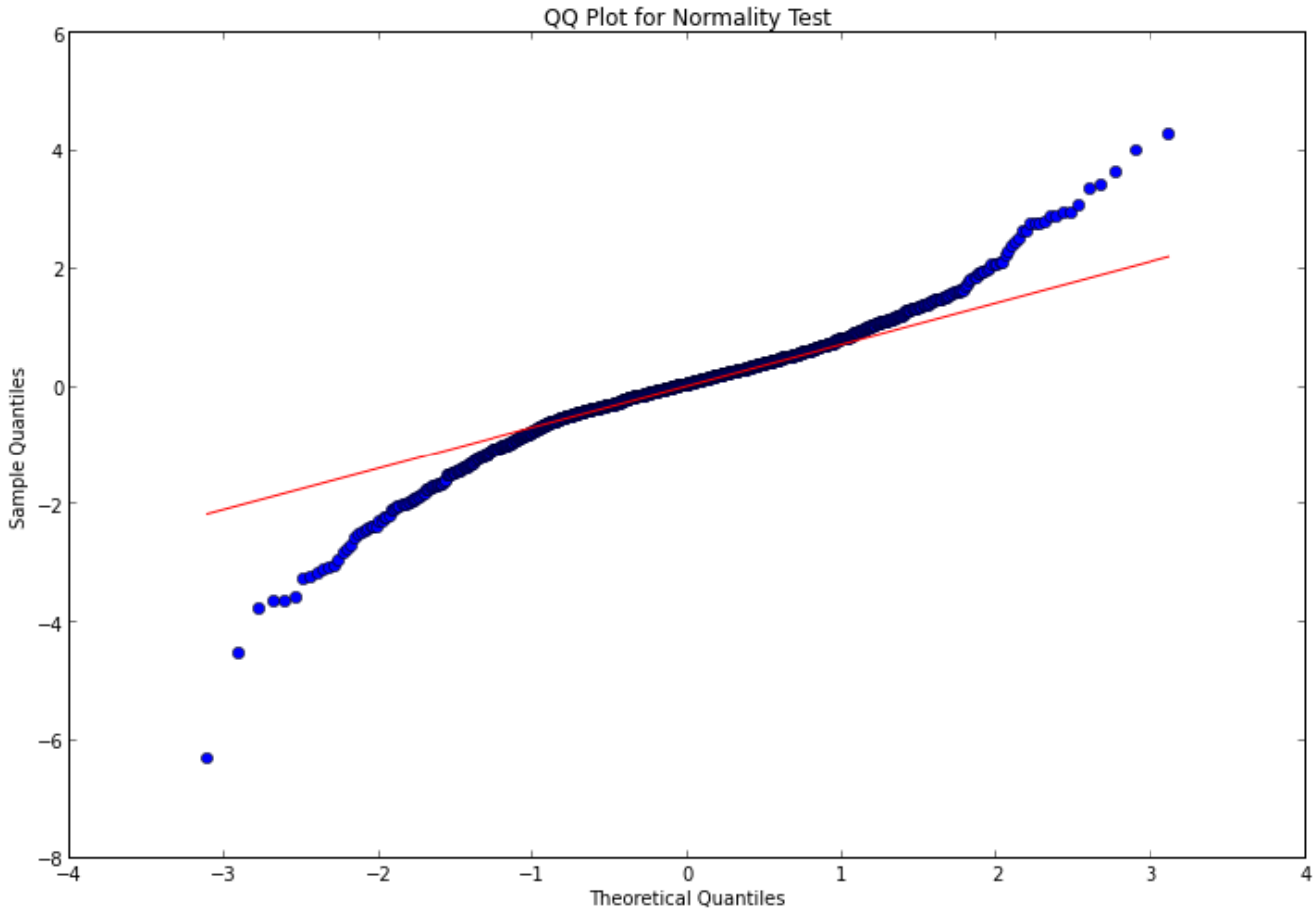

Figure 14: We do the normality test for residuals of ARIMA model.

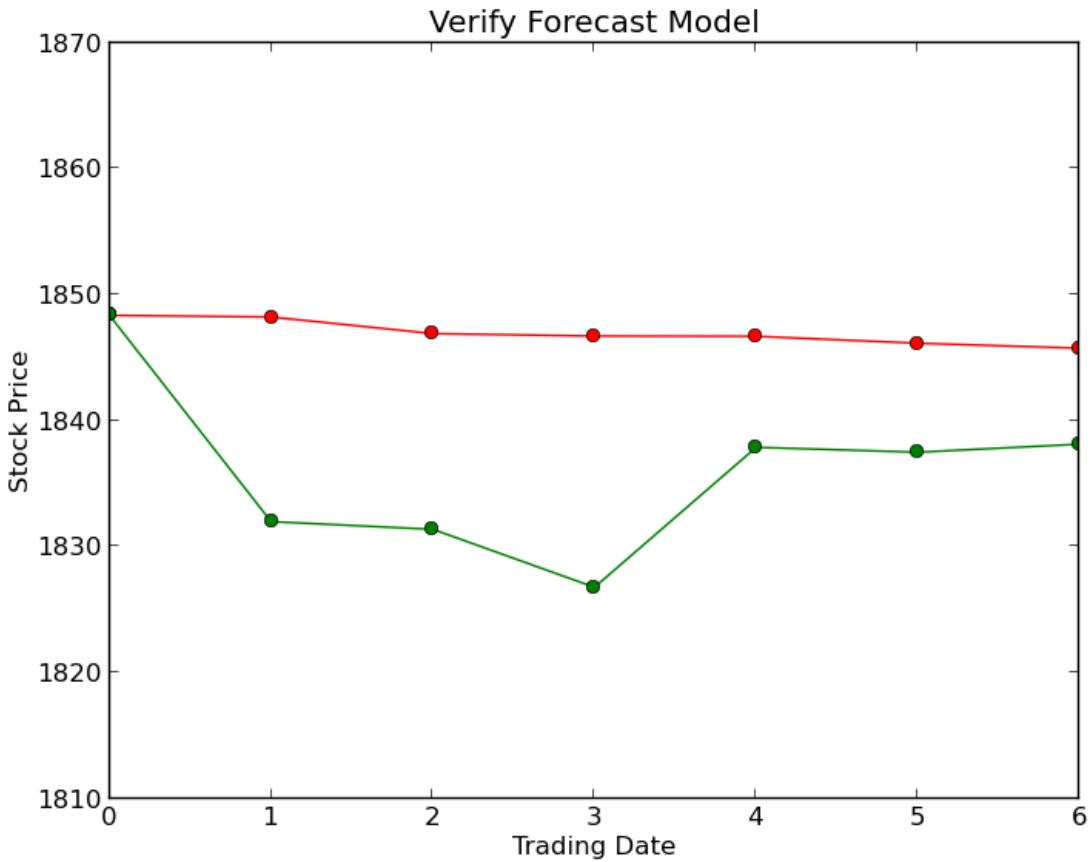

Figure 15: We used cross-validation again to verify this time series model. We chose the daily prices from 2010 to 2013 as training set and second week of 2014 as testing set and plotted the result. The red line is the predicted prices, and the green line is the actual prices.

Thank You

Thank you!