Long-term Stock Market Forecasting using Gaussian Processes

Anonymous Author(s)
Affiliation
Address
email

Abstract

Forecasting stock market prices is an attractive topic to researchers from different fields. The accuracy of this forecasting is very critical for market dealers. The existing forecast models show valid results in short-term forecasting; however, the accuracy of these models degrades in long-term forecasting. In this project, the Gaussian processes are applied to forecast the stock market trend. We select three stocks from NASDAQ Stock Market to test the proposed model. The experiment results show worthy findings of the stocks behavior over different periods of time. This model could help investors to make the long-term investment or to validate their investment decisions.

1 Introduction

Nowadays, most of the stock market traders relay on machine learning techniques to analyze and forecast stock prices and index changes. The accuracy of these techniques is still an issue due to several factors such as seasons, political situation and economic conditions that cause fluctuation of stock market movement (Ou & Wang, 2009). Although this movement does not follow exact seasonal cycles all the time, it is highly recommended not to ignore these cycles (Jeffrey & Kass, 2012). This project proposes a new application of long-term forecasting with the Gaussian processes (GP) model in (Chapados & Bengio, 2007) in stock market.

In general, there are two methodologies to predict stock prices: Fundamental Analysis and Technical Analysis. The Fundamental Analysis relies on the past performance of the company to make predictions. The Technical Analysis deals with past stock prices to understand its pattern change and predict the future prices. Although, most of machine learning application show more interest in Technical Analysis, hybrid approaches could combine both methodologies to make prediction (Ayodele, et al., 2012). In this paper, Technical Analysis will be used to perform long-term predictions in stock prices.

1.2 Motivation

In stock market, investors need long-term forecasting techniques to choose the right time to buy/sell stocks to maximize their profits or to minimize their loss. The majority of existing stock market forecasting techniques require predictions over a single continuous time series. These techniques perform well in short-term (a day to weeks) time series prediction but the accuracy of these techniques degrades when long-term time series prediction is made. The motivation for this project comes from the presence of large amount of historical data in stock market and the ability to use of GP in long-term time series forecasting (Chapados & Bengio, 2007). The goal of this project is to help investors to choose the right stock to invest in, based on long-term forecasting. Also, this project can assist investors to predict the right time to buy/sell in stock market to maximize the profit.

The rest of this paper is organized as follows. Section 2 sheds light on the related work and gives background about stock market and Gaussian processes GP .In Section 3, we present the methodology and the collected data. Section 4 gives a summary of the results obtained and the analysis of these results. Section 5 concludes with future directions of work.

2 Background and related work

Several forecasting models have considered Gaussian processes for time-series forecasting (Chapados & Bengio, 2007; Todd & Correa, 2007; Groot et al., 2011). In this section, we give an overview about some related studies. Also, brief introductions about stock market and Gaussian processes are covered.

Stock market trend prediction using Gaussian processes were tackled in (Todd & Correa, 2007). This study shows that increasing the size of training data (a long time period) gives more accurate prediction. The drawback of this approach is the high computational time. Multiple-step time series forecasting using sparse Gaussian process was addressed in (Groot et al., 2011). This approach produced more accurate and faster predictions than standard GP approach. Chapados and Bengio (2007), introduced a Long-term forecasting approach using Gaussian processes. This approach used functional representation of time series to perform long-term forecasting. Commodity spread trading data was used as an application for this approach. In this project, the technique in (Chapados & Bengio, 2007) is applied to forecast long-term prices in stock market.

2.1 Stock Market

Stock markets are public markets for trading the companies' stocks (shares) at agreed prices. Investors (companies or individuals) are allowed to buy and sell stocks and these transactions are called trading. The stock prices depend on the demands and supplies; it goes high when there is high demand and falls down at low demand. In stock market, a quarter (Q) refers to one-fourth of a year. The four quarters are: January, February and March (Q1); April, May and June (Q2); July, August and September (Q3); and October, November and December (Q4). Investors use the past several quarters to forecast the future of the stocks (Wikipedia, 2013).

Stock markets are considered as one of the economic indicators of countries. The growth of stock prices attracts investors and increases the companies' values. In general, the growth in stock market reflects the strength and development of countries' economics so that countries watch and control the behavior of stock market (Preethi & Santhi, 2012). The size of global stock market was estimated at about \$54 Trillion in 2010 (anonymous, 2012).

2.2 Gaussian Processes

A Gaussian process (GP) is a popular technique in machine learning and is widely used in time series analysis (Mori & Ohmi, 2005). Rasmussen and Williams (2006) defined GP as "a collection of random variables, any finite number of which have a joint Gaussian distribution". The GP is used to characterize probability distribution over functions by defining two functions: mean function $\mathrm{m}(\mathrm{x})$ and the covariance function mean function $\mathrm{k}(\mathrm{x}_1, \mathrm{x}_2)$ (Rasmussen & Nickisch, 2006). To describe a real process $\mathrm{f}(\mathrm{x})$ as a GP, we write:

$$\mathrm{f}(\mathrm{x}) \sim \mathcal{GP}\big(\mathrm{m}(\mathrm{x}), \mathrm{k}(\mathrm{x}_1, \mathrm{x}_2)\big), \tag{1}$$

where,

$$\mathrm{m}(\mathrm{x}) = \mathbb{E}[\mathrm{f}(\mathrm{x})],$$
$$\mathrm{k}(\mathrm{x}_1, \mathrm{x}_2) = \mathrm{E}\big[\big(\mathrm{f}(\mathrm{x}_1) - \mathrm{m}(\mathrm{x}_1)\big)\big(\mathrm{f}(\mathrm{x}_2) - \mathrm{m}(\mathrm{x}_2)\big)\big].$$

In regression, given a data set D of N observations; $\mathrm{D} = \{(\mathrm{x}_\mathrm{i}, \mathrm{y}_\mathrm{i}) \mid \mathrm{i} = 1, \dots, \mathrm{N}\}$, with $\mathrm{x}_\mathrm{i} \in \mathbb{R}^\mathrm{D}$

and $y_i \in \mathbb{R}$, the goal is to predict new y_* given x_* using $f(x)$ such that: $y_i = f(x_i) + \delta_i$ where δ_i is a Gaussian noise with mean zero and variance σ^2. However, we assume that closing prices in stock market are noise free because true prices are evaluated at closing time (Todd & Correa, 2007). The prior distribution of the observed target y is given by

$$y \sim \mathcal{N}(0, K(X, X)), \tag{2}$$

where, $K(X, X)$ is the covariance matrix between all pairs of training points and X is $(n\times m)$ matrix of input. In this project, (Gaussian) radial basis function kernel, or RBF kernel is used:

$$k(x_1, x_2) = \exp\left(-\sigma \|x_1 - x_2\|^2\right). \tag{3}$$

The predictive distribution of y_* can be computed by conditioning on the training data to get $p(f(x_*)|x_*, D)$. The joint distribution over y and predictions of x_* is given by:

$$\begin{bmatrix} y \\ f(x_*) \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(X, X) & K(X, x_*) \\ K(x_*, X) & K(x_*, x_*) \end{bmatrix}\right). \tag{4}$$

The conditional distribution of (2) allows us to get the predictive distribution of y_* with the following mean and covariance (Ou & Wang, 2011):

$$\bar{f}(x_*) = K(X, x_*)^T (K + \sigma_n^2 I)^{-1} y, \tag{5}$$

$$V_f(x_*) = K(x_*, x_*) - K(X, x_*)^T (K + \sigma_n^2 I)^{-1} K(X, x_*) \quad . \tag{6}$$

3 Methodology

The main idea of this approach is to avoid representing the whole history as one time series. Each time series is treated as an independent input variable in the regression model (Chapados & Bengio, 2007). For trading year i, there are M_i trading days, $i = 1, \ldots, N$ and $t = 1, \ldots, M_i$. The model problem is given M observations from $i = 1, \ldots, N-1$ trading years and partial trading days from N, $\{y_t^N\}$, $t = 1, \ldots, M_N$, we want to predict the rest of trading days in N, $\{y_\tau^N\}$, $\tau = M_N + 1, \ldots, M_N + H$, where $M_N + H$ is the last day of trading in N. Also, it is given $\{x_t^i\}$ for each series and our objective is to find $P(\{y_\tau^N\}, M_N + 1, \ldots, M_N + H | \{x_t^i, y_t^i\}_{t=1,\ldots,M_i}^{i=1,\ldots,N})$. See Figure 1.

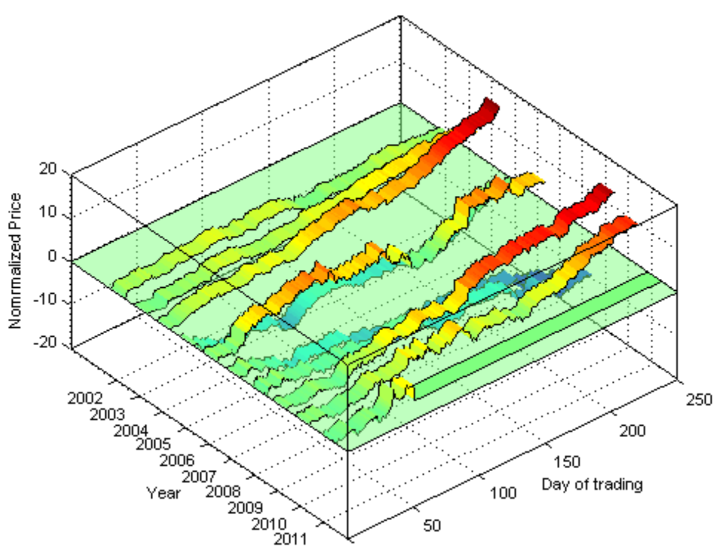

Figure 1: Illustration of the regression variables (price history from 2002 to the first quarter of 2011) of Starbucks stock. The objective of this model is to predict the "green strip" in 2011.

3.1 Data description

For this project, three random stocks were randomly selected from NASDAQ Stock Market, namely Hewlett-Packard Company (HPQ), Yahoo Inc. (YHOO) and Starbucks Corporation (SBUX). The daily changes of closing prices of these stocks were examined. The historical data was downloaded from the yahoo finance section.

The sample period is from Jan 01 2002 to Dec. 31 2011 (N = 10). We have about 250 days of trading per year since no data is observed on weekends. However, some years have more than 250 days of trading ($M_i = 250$), we choose to ignore these days so that the whole sample is of 2500 trading days.

We choose to use adjusted close prices because we aim to predict the trend of the stocks not the prices. The adjusted close price is used to avoid the effect of dividends and splits because when stock has a split, its price drop by half.

The adjusted close prices are standardized to zero mean and unit standard deviation. We also normalize the prices in each year to avoid the variation from previous years by subtracting the first day to start from zero.

As time-series model, we include a representation of the trading date as independent (input) variables. The trading date is split into two parts: the trading year i (an integer, from 1 to 10) and the days of trading t (an integer, from 1 to 250), Figure 1. These variables are preprocessed before using them as input to the GP. They were standardized to zero mean and unit standard deviation.

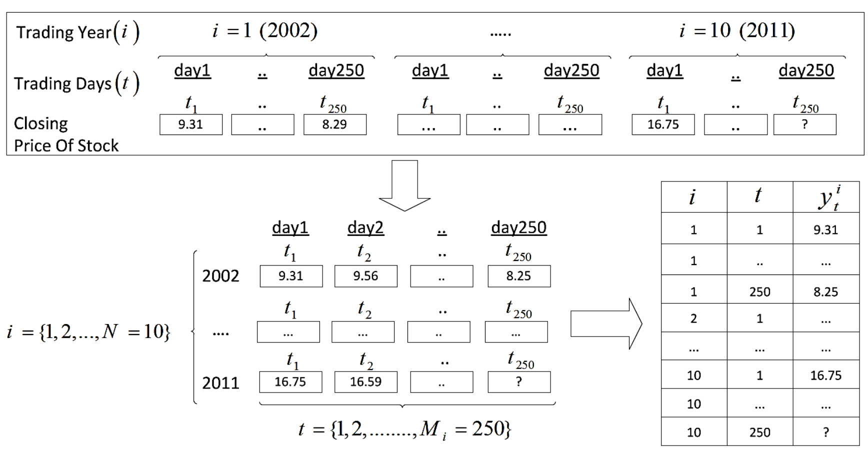

Figure 2: Example of data processing to split trading date into two inputs: trading year (***i***) and trading day (***t***)

4 Evaluation

To evaluate the performance of the proposed approach, "kernlab" R package is used. For each stock, we applied two scenarios for long-term forecasting. The first scenario, given complete observations from 9 years (2002 to 2010) and the first quarter (Q1) from 2011, we want to predict the second, third and fourth quarters of 2011 (Q2, Q3 and Q4). The data is divided into two sub-samples where the training data spans from Jan 01 2002 to the first quarter of 2011 with 2312 trading days. The rest trading days of year 2011 of size 188 days are reserved for test data.

The second scenario, given complete observations from 9 years (2002 to 2010) and the first and second quarters (Q1 and Q2) from 2011, we want to predict the third and fourth quarters of 2011 (Q3 and Q4). The data is divided into two sub-samples where the training data spans from Jan 01 2002 to the second quarter of 2011 with 2374 trading days. The rest trading days of year 2011 of size 126 days are reserved for test data. Figure 3 shows the training data and the forecast results for Starbucks stock.

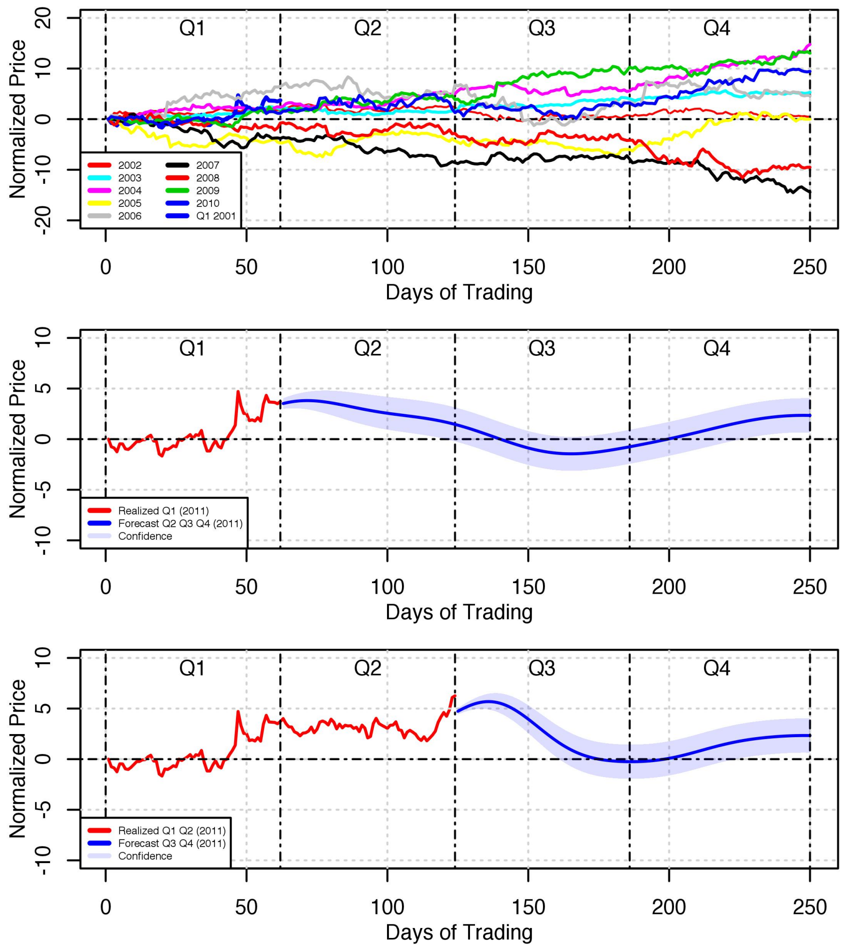

Figure 3: Top plot: Training set of Starbucks stock for the period from 2002 to the first quarter of 2011. Each line represent complete trading year. Meddle plot: Shows the first scenario where forecast made for the rest quarters of 2011 (Q2, Q3 and Q4). Bottom plot: shows the second where training set is the period from 2002 to the second quarter of 2011. Forecast made for the third and fourth quarters of 2011.

4.1 Results and discussion

The forecast results for the three stocks (HP, Yahoo and Starbucks) are shown in Figure 4, 5, 6. The "blue" lines show the forecast prices and the "black" lines show the actual prices. In Figure 4, the results of scenario 1 (top part) shows drop in HP stock prices in Q2, Q3 and Q4 of 2011. Also, scenario 2 (bottom part) confirms this drop until the end of 2011. Based on that, investors should not buy HP stock in 2011 and if they already did, it is highly recommended to sell it to minimize their loss. Although, the model could not predict the high drop in Q3, it keeps following the trend of the actual prices.

Figure 5 shows the forecast price of Yahoo stock. The results of scenario 1 (top part) show slight decrease in Yahoo stock prices in Q2 and Q3 of 2011; however, the price shows some improvement in Q4. The second scenario shows Yahoo stock prices reverse direction in Q4. Investors can take the risk and buy in Q3 or wait until the beginning of Q4. The forecasting model is able to track the trend of this stock most of the time.

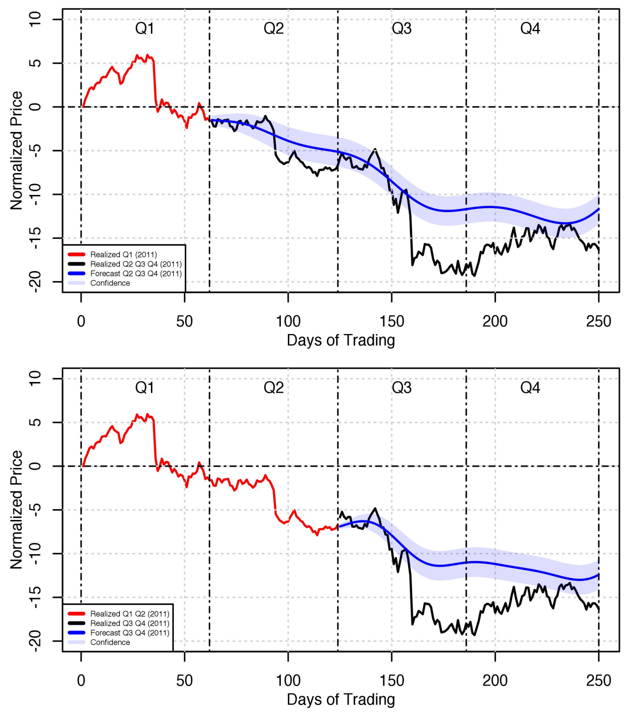

Figure 4: Top part: Forecast result for HP stock from scenario 1. Bottom part: Forecast result for HP stock from scenario 2.

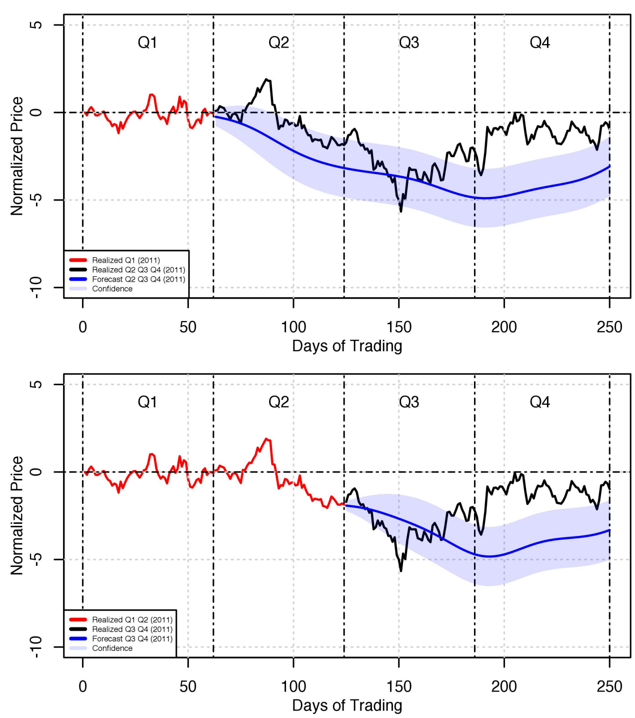

Figure 5: Top part: Forecast result for Yahoo stock from scenario 1. Bottom part: Forecast result for Yahoo stock from scenario 2.

The forecasting result for Starbucks stock is sown in Figure 6. Although, the true model shows high fluctuation in 2011, our model keeps following the main trend of the stock. Scenario 1 shows falling in the price until the mid of Q3, however, scenario 2 updates the curve in Q3 to follow the increase at the end of Q2. Both scenarios agree that the mid of Q3 is suitable to buy this stock. If investors own the stock before Q3, it is highly recommended to wait until the end of Q4.

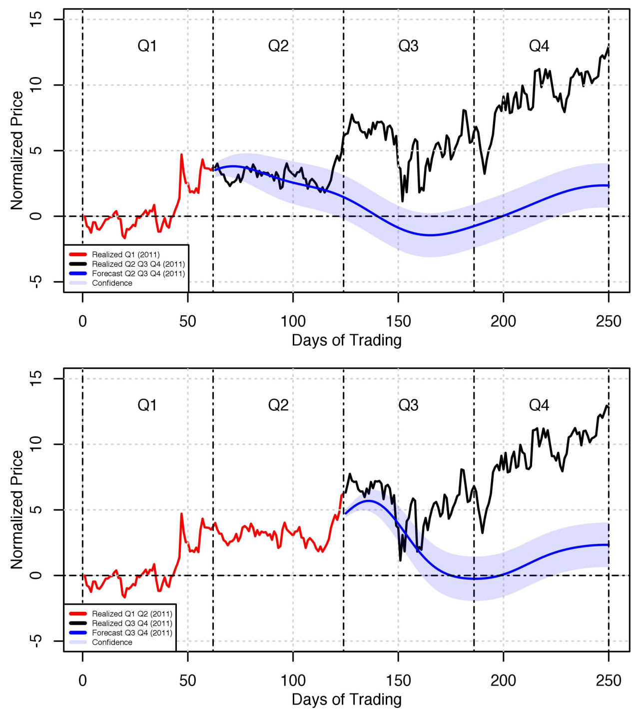

Figure 6: Top part: Forecast result for Starbucks stock from scenario 1. Bottom part: Forecast result for Starbucks stock from scenario 2.

In general, this model is able to track the prices of the three stocks. As we know, stock price could be affected by several factors such as political situation and economic conditions, which may cause high fluctuations as shown in some areas of this experiment. As a long-term forecasting model, it is acceptable to not follow these fluctuations.

5 Conclusion and future work

In this project, we applied Gaussian processes to perform long-term forecasting in stock market. This technique showed acceptable prediction to three stocks from NASDAQ Stock Market. The experiment showed highly acceptable time to buy and sell over different period of times. Due to the fast computation and the simplicity of this model, investors could use this model to do a long-term investment or to validate their investment decisions. More stocks could be tested on this model from other stock market.

References

Ayodele, A., Charles, A., Marion, A. & Otokiti Sunday O. (2012). "Stock Price Prediction using Neural Network with Hybridized Market Indicators. *Journal of Emerging Trends in Computing and*

Information Sciences, VOL. 3: 1, 1-9.

Chapados, N. & Bengio, Y. (2007). Forecasting Commodity Contract Spreads with Gaussian Process, *in 13th International Conference on Computing in Economics and Finance*, June 14 - 16, 2007, Montréal, Quebec, Canada.

Groot, P., Lucas, P. & Paul van den Bosch. (2011). Multiple-step Time Series Forecasting with Sparse Gaussian Processes, *BNAIC*, 1-8.

Jeffrey, A. & Kass, D. (2012). The Little Book of Stock Market Cycles (Little Books. Big Profits), Wiley.

Mori, H. & Ohmi M. (2005).Probabilistic short-term load forecasting with Gaussian processes. *Proceedings of the 13th International Conference on Intelligent Systems Application to Power System (ISAP)*, November 6-10, 2005, Arlington, Virginia, 452-457.

Ou, P. & Wang, H. (2009). Prediction of market index movement by ten data mining techniques. *Modern Applied Science*, 3:12, 28–42.

Ou, P. & Wang, H. (2011). Modeling and Forecasting Stock Market Volatility by Gaussian Processes based on GARCH, EGARCH and GJR Models. *Proceedings of the World Congress on Engineering*, July 6-8, 2011, London, U.K., 338-342.

Preethi, G. & Santhi, B. (2012).Stock market forecasting techniques: a survey. *Journal of Theoretical and Applied Information Technology*, 46:1, 24-30.

Rasmussen, C. & Nickisch, H. (2006). Gaussian Processes for Machine Learning (GPML) Toolbox. *Journal of Machine Learning Research*, 11, 3011-3015.

Todd, M. & Correa, A. (2007). Gaussian Process Regression Models for Predicting Stock Trends. Technical Report on MIT University.

Wikipedia: The free encyclopedia. (2013) Stock market. Retrieved April 2, 2013, from http://en.wikipedia.org/wiki/Stock_market.

World Capital Markets – Size of Global Stock and Bond Markets. Retrieved April 1, 2013, from http://qvmgroup.com/invest/2012/04/02/world-capital-markets-size-of-global-stock-and-bond-markets/.