

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 rely 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

43 in, based on long-term forecasting. Also, this project can assist investors to predict the right
44 time to buy/sell in stock market to maximize the profit.

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

49 50 **2 Background and related work**

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

55 Stock market trend prediction using Gaussian processes were tackled in (Todd & Correa,
56 2007). This study shows that increasing the size of training data (a long time period) gives
57 more accurate prediction. The drawback of this approach is the high computational time.
58 Multiple-step time series forecasting using sparse Gaussian process was addressed in (Groot
59 et al., 2011). This approach produced more accurate and faster predictions than standard GP
60 approach. Chapados and Bengio (2007), introduced a Long-term forecasting approach using
61 Gaussian processes. This approach used functional representation of time series to perform
62 long-term forecasting. Commodity spread trading data was used as an application for this
63 approach. In this project, the technique in (Chapados & Bengio, 2007) is applied to forecast
64 long-term prices in stock market.

65 66 **2.1 Stock Market**

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

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

80 81 **2.2 Gaussian Processes**

82 A Gaussian process (GP) is a popular technique in machine learning and is widely used in
83 time series analysis (Mori & Ohmi, 2005). Rasmussen and Williams (2006) defined GP as "a
84 collection of random variables, any finite number of which have a joint Gaussian
85 distribution". The GP is used to characterize probability distribution over functions by
86 defining two functions: mean function $m(x)$ and the covariance function mean function
87 $k(x_1, x_2)$ (Rasmussen & Nickisch, 2006). To describe a real process $f(x)$ as a GP, we write:

$$88 f(x) \sim \mathcal{GP}(m(x), k(x_1, x_2)), \quad (1)$$

89 where,

$$90 m(x) = \mathbb{E}[f(x)], \\ k(x_1, x_2) = E[(f(x_1) - m(x_1))(f(x_2) - m(x_2))].$$

91 In regression, given a data set D of N observations; $D = \{(x_i, y_i) | i = 1, \dots, N\}$, with $x_i \in \mathbb{R}^D$

92 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
 93 δ_i is a Gaussian noise with mean zero and variance σ^2 . However, we assume that closing
 94 prices in stock market are noise free because true prices are evaluated at closing time (Todd
 95 & Correa, 2007). The prior distribution of the observed target y is given by

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

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

$$k(x_1, x_2) = \exp(-\sigma \|x_1 - x_2\|^2). \quad (3)$$

101 The predictive distribution of y_* can be computed by conditioning on the training data to get
 102 $p(f(x_*)|x_*, D)$. The joint distribution over y and predictions of x_* is given by:

$$103 \begin{bmatrix} y \\ f(x_*) \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(X, X) & K(X, x_*) \\ K(x_*, X) & K(x_*, x_*) \end{bmatrix}\right). \quad (4)$$

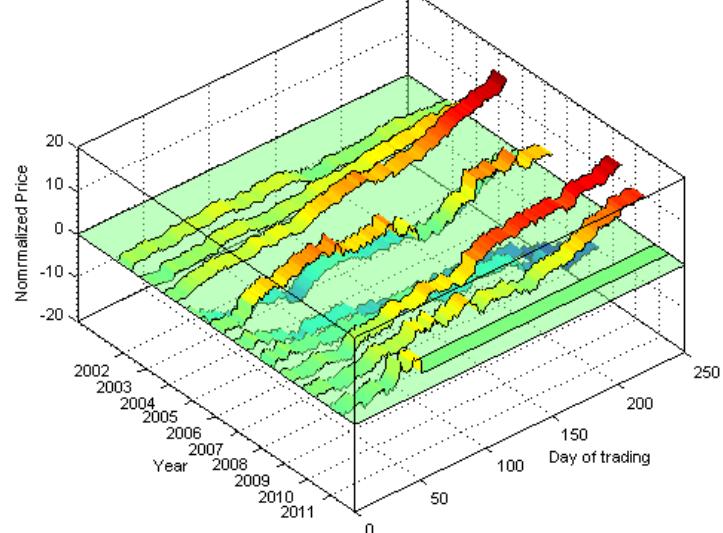
104 The conditional distribution of (2) allows us to get the predictive distribution of y_* with the
 105 following mean and covariance (Ou & Wang, 2011):

$$106 \bar{f}(x_*) = K(X, x_*)^T(K + \sigma_n^2 I)^{-1}y, \quad (5)$$

$$107 V_f(x_*) = K(x_*, x_*) - K(X, x_*)^T(K + \sigma_n^2 I)^{-1}K(X, x_*) \quad . \quad (6)$$

109 3 Methodology

110 The main idea of this approach is to avoid representing the whole history as one time series.
 111 Each time series is treated as an independent input variable in the regression model
 112 (Chapados & Bengio, 2007). For trading year i , there are M_i trading days, $i = 1, \dots, N$ and
 113 $t = 1, \dots, M_i$. The model problem is given M observations from $i = 1, \dots, N - 1$ trading years
 114 and partial trading days from N , $\{y_t^N\}$, $t = 1, \dots, M_N$, we want to predict the rest of trading
 115 days in N , $\{y_\tau^N\}$, $\tau = M_N + 1, \dots, M_N + H$, where $M_N + H$ is the last day of trading in N . Also,
 116 it is given $\{x_t^i\}$ for each series and our objective is to find $P(\{y_\tau^N\}, M_N + 1, \dots, M_N +$
 117 $H | \{x_t^i, y_t^i\}_{t=1, \dots, M_i}^i)$. See Figure 1.



119
 120 Figure 1: Illustration of the regression variables (price history from 2002 to the first quarter
 121 of 2011) of Starbucks stock. The objective of this model is to predict the "green strip" in
 122 2011.

123 **3.1 Data description**

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

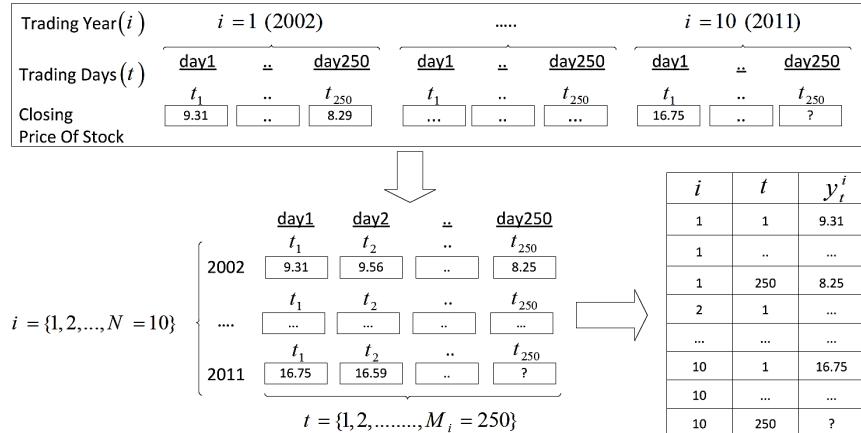
128 The sample period is from Jan 01 2002 to Dec. 31 2011 ($N = 10$). We have about 250 days
 129 of trading per year since no data is observed on weekends. However, some years have more
 130 than 250 days of trading ($M_i = 250$), we choose to ignore these days so that the whole
 131 sample is of 2500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

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

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

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

143



144

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

147

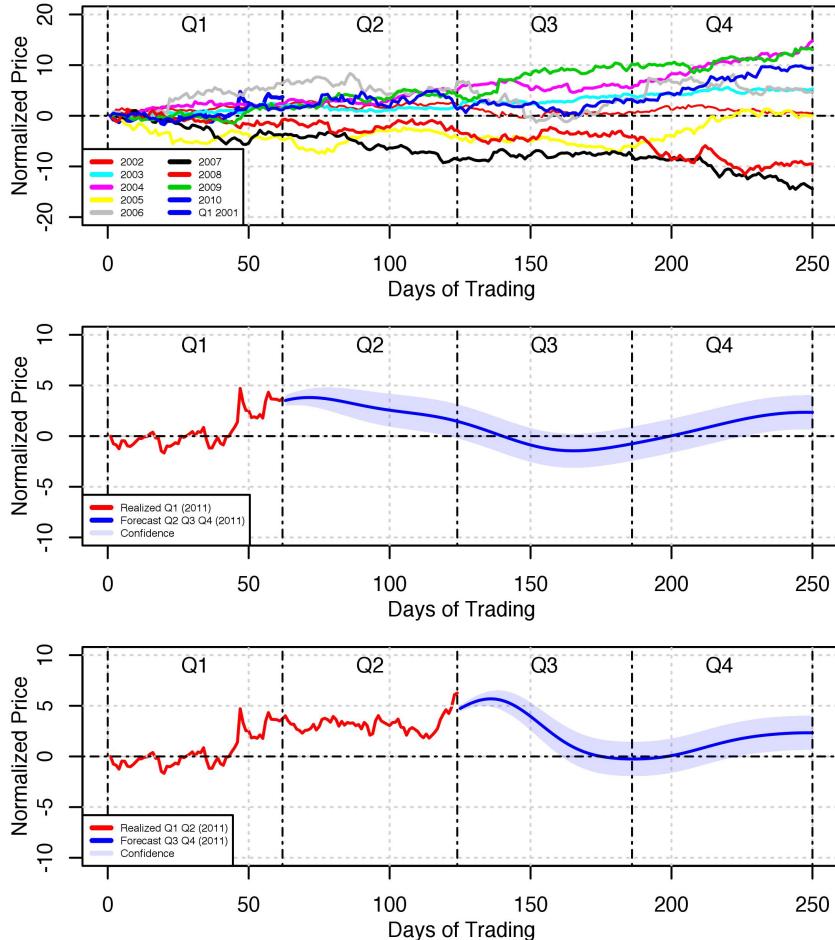
148 **4 Evaluation**

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

156 The second scenario, given complete observations from 9 years (2002 to 2010) and the first
 157 and second quarters (Q1 and Q2) from 2011, we want to predict the third and fourth quarters
 158 of 2011 (Q3 and Q4). The data is divided into two sub-samples where the training data spans
 159 from Jan 01 2002 to the second quarter of 2011 with 2374 trading days. The rest trading

160 days of year 2011 of size 126 days are reserved for test data. Figure 3 shows the training
161 data and the forecast results for Starbucks stock.

162



163

164 Figure 3: Top plot: Training set of Starbucks stock for the period from 2002 to the first quarter of
165 2011. Each line represent complete trading year. Meddle plot: Shows the first scenario where
166 forecast made for the rest quarters of 2011 (Q2, Q3 and Q4). Bottom plot: shows the second where
167 training set is the period from 2002 to the second quarter of 2011. Forecast made for the third and
168 fourth quarters of 2011.

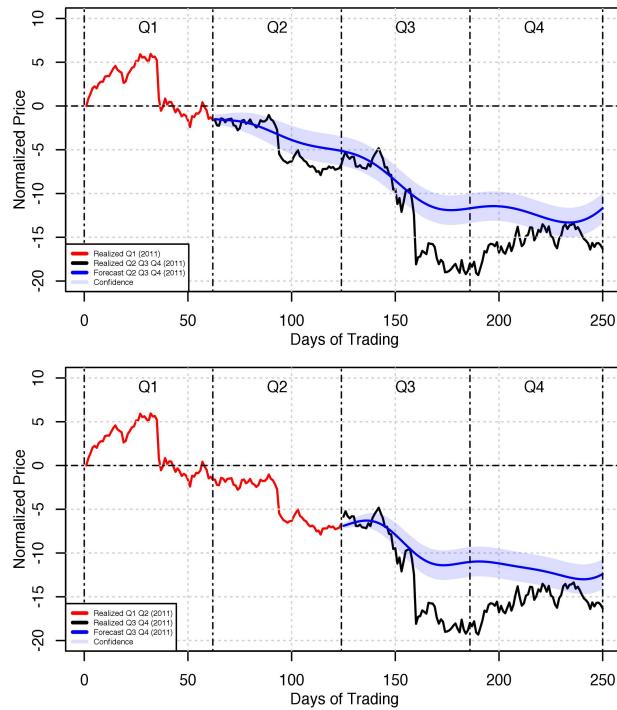
169

170 4.1 Results and discussion

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

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

183

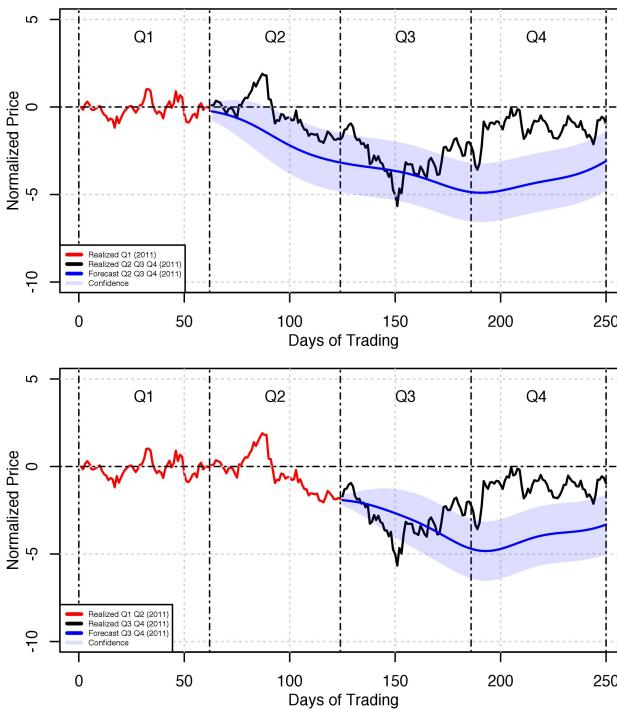


184

185
186

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

187



188

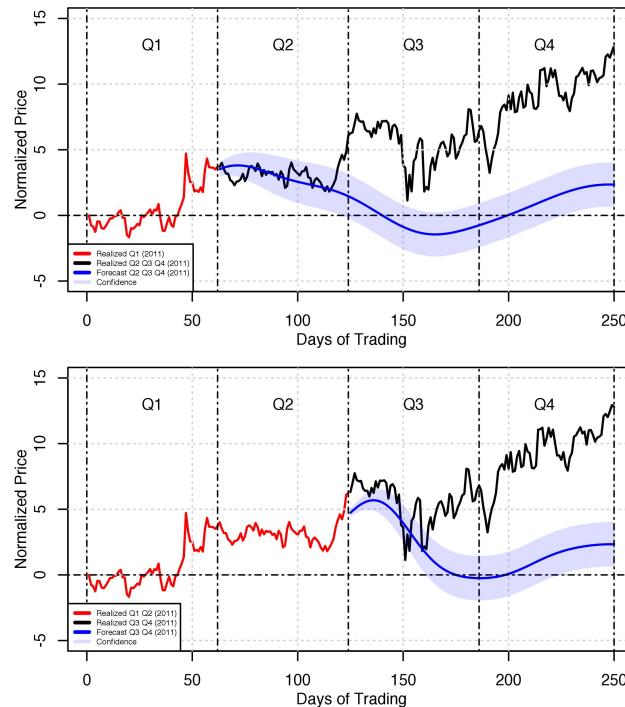
189
190

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

191

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

198



199

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

202

203 In general, this model is able to track the prices of the three stocks. As we know, stock price
204 could be affected by several factors such as political situation and economic conditions,
205 which may cause high fluctuations as shown in some areas of this experiment. As a long-
206 term forecasting model, it is acceptable to not follow these fluctuations.
207

208 **5 Conclusion and future work**

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

215 **References**

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

- 218 *Information Sciences*, VOL. 3: 1, 1-9.
- 219 Chapados, N. & Bengio, Y. (2007). Forecasting Commodity Contract Spreads with Gaussian Process,
220 *in 13th International Conference on Computing in Economics and Finance*, June 14 - 16, 2007,
221 Montréal, Quebec, Canada.
- 222 Groot, P., Lucas, P. & Paul van den Bosch. (2011). Multiple-step Time Series Forecasting with Sparse
223 Gaussian Processes, *BNAIC*, 1-8.
- 224 Jeffrey, A. & Kass, D. (2012). The Little Book of Stock Market Cycles (Little Books. Big Profits),
225 Wiley.
- 226 Mori, H. & Ohmi M. (2005).Probabilistic short-term load forecasting with Gaussian processes.
227 *Proceedings of the 13th International Conference on Intelligent Systems Application to Power System*
228 (*ISAP*), November 6-10, 2005, Arlington, Virginia, 452-457.
- 229 Ou, P. & Wang, H. (2009). Prediction of market index movement by ten data mining techniques.
230 *Modern Applied Science*, 3:12, 28-42.
- 231 Ou, P. & Wang, H. (2011). Modeling and Forecasting Stock Market Volatility by Gaussian Processes
232 based on GARCH, EGARCH and GJR Models. *Proceedings of the World Congress on Engineering*,
233 July 6-8, 2011, London, U.K., 338-342.
- 234 Preethi, G. & Santhi, B. (2012).Stock market forecasting techniques: a survey. *Journal of Theoretical*
235 *and Applied Information Technology*, 46:1, 24-30.
- 236 Rasmussen, C. & Nickisch, H. (2006). Gaussian Processes for Machine Learning (GPML) Toolbox.
237 *Journal of Machine Learning Research*, 11, 3011-3015.
- 238 Todd, M. & Correa, A. (2007). Gaussian Process Regression Models for Predicting Stock Trends.
239 Technical Report on MIT University.
- 240 Wikipedia: The free encyclopedia. (2013) Stock market. Retrieved April 2, 2013, from
241 http://en.wikipedia.org/wiki/Stock_market.
- 242 World Capital Markets – Size of Global Stock and Bond Markets. Retrieved April 1, 2013, from
243 <http://qvmgroup.com/invest/2012/04/02/world-capital-markets-size-of-global-stock-and-bond-markets/>.
- 245