Stock Market Prediction using Digital Signal Processing Models

Shashanklyer
Student (ELEX)
D.J.Sanghvi College of Engg,
Plot No. U-15, JVPD Scheme,
Bhaktivedanta Swami Marg,
Vile Parle (W), Mumbai-400056

Nisarg R. Kamdar
Student (ELEX)
D.J.Sanghvi College of Engg,
Plot No. U-15, JVPD Scheme,
Bhaktivedanta Swami Marg,
Vile Parle (W), Mumbai-400056

BaharSoparkar
Assistant Professor (ELEX)
D.J.Sanghvi College of Engg,
Plot No. U-15, JVPD Scheme,
Bhaktivedanta Swami Marg,
Vile Parle(W), Mumbai-400056

ABSTRACT
This paper aims to exploit the temporal correlation that exists between the various stock market variablesemploying concepts of adaptive filters and signal modelling in order to predict future trends and prices, using two statistical processes. Linear regression algorithm (Gradient Descent)has been used for real time prediction. The Finite Impulse Response (FIR) adaptive filter is an iterative process that minimizes the mean square error. An extrapolation of Prony's Normal Equation has been used to predict values using the least square estimation. This models cross-section regression, i.e. the relationship between variables at a particular point in time. The analysis has been performed on stocks listed on NSADAQ and the mean square error was compared. This study reveals that the DSP techniques are adequate for modeling thevariation in stock prices.

General Terms
Prediction techniques, Digital Signal Processing, Statistical Signal Modeling, Power Spectral Density

Keywords
Stock Market Prediction, DSP, Statistical Signal Processing, Regression models, Prony's Algorithm

1. INTRODUCTION
Prediction of stock indices has always been an enticing and intriguing field.Predicting the stock market behavior through techniques and various methods is a useful tool to assist investors to act with greater certainty, and taking the risks, and volatility of an investment into consideration and know when to buy the cheapest price and when to sell to highest price. [1]Increase and decrease of stock marketprices depends on various factors such as amount of demand, exchange rate, price of gold, price of oil, political and economic events, but in the other view point we canconsider the stock market price variation as time series andwithout notation to the mentioned factors, and just by findingthe sequence rules of price train, make the price prediction inthe future. [2]The aim here is to achieve a fair degree of accuracy using two different prediction models and then using the mean squared error to provide an unbiased comparison of the results.The variable being predicted istoday's closing price and opening price using either the previous day's opening, low, high and adjusted close price or using a time series of the closing prices.The first step is to acquire large amounts of historical data for analysis. The concept of data scraping has been used.The first algorithm is the Gradient descent/Steepest descent algorithm. This linear regression algorithm aims to generate a hypothesis function that will closely approximate the correct output signal.Linear regression is in essence curve fitting. Using the set of input training data, a curve is plotted, the equation of which is the hypothesis function. The cost function is the mean squared error between the predicted output and the actual output. The minimized cost function aids in weight updation and improvement of the hypothesis function. The second algorithm is an interpretation of Prony's Normal Equation algorithm that can compute the solution in a single step while gradient descent would require numerous iterations. The use of such all pole systems has a computational advantage over the more general pole-zero models. [3] The poles of are used to construct an augmented Normal Equation. The poles are found such that they minimize the squared error. [4]This model is particularly suited for short term price estimation.

2. LITERATURE REVIEW
Since the 1970s researchers from the financial services industry have been employing a variety to technical analysis tools to identify trends in the market.Fama and French (1992) attempted to build over the Capital Asset Pricing model by linking the movement of stock returns to the three fundamental variables, namely earnings yield, book-to-market ratio, and size. The model is based on the authors' research which suggested that value stocks outperform growth stocks and small capital outperform large capital stocks.Altman [1968] was the pioneer the development of logistic regression, which is often used when outcome to be predicted is binary in nature. Ohlson [1980] later used LR to construct the defaultprediction model. [5]An experiment which was conducted by Olaniyi, Adewole&Jimoh which used linear regression line to generate new knowledge from historical data, and identified the patterns that describe the stock trend (Olaniyi, Adewole&Jimoh, 2011) [6]The authors sounded an optimistic note regarding the inherent ability of linear regression to capture correlations and project them into the future. A bouquet of factors, such as political stability (measured by a proxy Worldwide Governance Index), strength of corporate management and financial ratios are variables that have been considered for stock price predictions (Ou, P. and Wang, H., 2009; Fama and French, 1993; Cochrane, 1988; Campel, 1987; Chen. et.al.1986).Quite surprisingly the focused use of stock price data was minimal. While financial ratios such as Earnings per Share, Price/Earnings Ratio, etc. are good indicators of the core fundamentals of a stock, their ability to model real time stock movements in uncertain and imperfect markets could be questioned. This anomaly provoked our study to understand the predictive quality of stock price data.

3. DATA TRANCHE

To obtain real time stock price data from the internet, data scraping has been used. The Python API opens the URL which includes the start and end date between which data points are to be retrieved. Yahoo! Finance website was used for this purpose. This data is pulled and sorted into an excel sheet. The excel sheet is converted into csv file which is accessed by algorithms. The excel sheet is refreshed to pull new data.Since the paper espouses two different approaches, to offer reasonable scope for comparison we decided to work with a single stock across both algorithms. For the analysis, the stock of Microsoft Corporation (NASDAQ:MSFT) was chosen. Our reasons for choosing this particular stock were multifold. Most importantly, Microsoft stock unlike Google or Apple stock hadn't undergone a split February 2003. This offered a fairly large tranche of stock data, without the stock value varying absurdly due to stock splits. Secondly Microsoft has had a fairly stable corporate governance structure in the period chosen for our analysis. This allows for the analysis to be strictly focused on market factors. And finally, Microsoft being a household brand name allows for better communication of our work. Data concerning the Microsoft stock between 9th June 2010 and 8th June 2015- a period of five years- was used. The choice of the period was deliberate. It is large enough to strongly depict and verify the predictive quality of the algorithms employed. Secondly it encompasses periods of extremely high volatility and bear markets (2011, 2012). The data tranche was divided into two parts for all algorithms. Two-thirds of it was used for training, while one-thirds was reserved for testing.

4. ALGORITHMS

4.1 Multiple Linear Regression

4.1.1 Principle

Regression predicts a numerical value. [7]Regression performs operations on a dataset where the target values have been defined already. And the result can be extended by adding new information.[8]The relations which regression establishes between predictor and target values can make a pattern.[9] This pattern can be used on other datasets which their target values are not known.[9]Linear regression is the firstmodel implemented. For more than one explanatory variable, the process is called multiple linear regression. (This term should be distinguished from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.)[10] The linear regression algorithms aim to generate a hypothesis function that will closely approximate the correct output value.The cost function is the vectorised multiplication of the weight vector by the vector of independent input variables given as:

i. $h(x) = \sum_{i=0}^{n} \theta_i x_i = \theta^T x$

Gradient descent is used to update the weights associated with the independent variables after the cost function is calculated.The cost function is the mean squared error between the predicted output and the actual output which is as follows, for 'm' number of input training samples

ii. $J(\theta_{(j)}) = \frac{1}{2}\left(\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2\right)$

This algorithm aims to minimize the cost function i.e. assuming that the cost function cure is bowl shaped, it moves towards the base of the bowl. The direction of steepest descent at any point in the plane is the direction that a marble would take if it were placed on the inside of this quadratic bowl. [11]

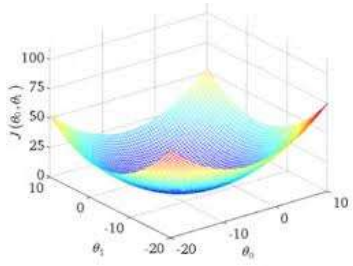

Fig 1. A quadratic bowl of two weights

The weight updation is performed by moving in the direction of the maximum descent, which is in the negative gradient direction.

iii. $$\frac{\partial}{\partial\theta_j}J(\theta) = \frac{\partial}{\partial\theta_j}\frac{1}{2}(h_\theta(x) - y)^2$$
$$= 2\cdot\frac{1}{2}(h_\theta(x) - y)\cdot\frac{\partial}{\partial\theta_j}(h_\theta(x) - y)$$
$$= (h_\theta(x) - y)\cdot\frac{\partial}{\partial\theta_j}\left(\sum_{i=0}^{n}\theta_i x_i - y\right)$$
$$= (h_\theta(x) - y)x_j$$

iv. $\theta_j := \theta_j + \alpha\left(y^{(i)} - h_\theta(x^i)\right)x_j^{(i)}$

4.1.2 Implementation

The weights are initialized to random values that are not the same. Having same initial values will cause the correlation between the input variables to have the same effect on the output.The independent variables are the previous day's open, low, high and adjusted closing price and the dependent variable is the today's closing price. The algorithm predicts the output which is divergent from the actual value and in subsequent steps betters this predicted output. The hypothesis function in this case will be the today's close price. Once the weights have been determined using the training data set, they should be used for testing in order to check their accuracy.

4.2 Normal Equation method

4.2.1 Principle

This algorithm implements regression using an m×n normal equation vector, where m is the number of training examples and n is the number of independent variables. Using an extrapolation of Prony's normal equation in the vectorised form, given by:

v. $\theta = (X^T X)^{-1}X^T\vec{y}$

where X is the vector of independent variables and Y is the vector of independent variables.[12]

Unlike gradient descent, which requires multiple iterations for convergance, this method uses the least squares estimation to calculate the weights in a single step. What it actually means is finding the values of θ_j that cause the partial derivative of the cost function with respect to θ_j to become equal to zero:

vi. $\frac{\partial}{\partial\theta_j}J(\theta) = 0$

There is also no need to set a value for the learning rate α. The issue that may arise is if the matrix $X^T X$ is non-invertible. This may happen if one or more independent variables is

actually dependent on the other or there are too many independent variables and not enough training examples $m \leq n$.

4.2.2 Implementation

The set of independent variables are the previous day's open, low, high, adjusted close and close price. These variables together for a given day are specified by a n×1vector given by :

x^j where j can have values 0 to m.

These individual vectors are amalgamated into a single vector X called the design matrix as:

vii. $$X = \begin{bmatrix} -(x^{(1)})^T - \\ -(x^{(2)})^T - \\ \vdots \\ -(x^{(m)})^T - \end{bmatrix}$$

Y is a 1 x m dimensional vector containing all the target values (today's closing prices corresponding to the rowx^j)and is given by:

viii. $$\vec{y} = \begin{bmatrix} y^1 \\ y^2 \\ \vdots \\ y^m \end{bmatrix}$$

Thus the weight matrix is calculated. These weights can be used to calculate the hypothesis function just like in gradient descent.

5. RESULTS

Goyal and Welch (2006) conducted a comprehensive examination of the existing evidence, and show that stock returns are hardly predictable in the out-of-sample context.[13] And hence the choice of the data set, as explained above, is the critical context within which the described results must be interpreted.To allow for definitive comparison between models, we calculate the mean squared error.

All graphs included in this section involve the plotting of actual values (Blue) against predicted values (green).

5.1 Linear Regression Results

The graph below is that of multivariable linear regression using gradient descent. It involves using previous day's opening price, high price, low price and adjusted closing price as independent variable to predict the dependent variable i.e. today's closing price.

Fig 2. Linear Regression-using multiple independent variables

This is the most accurate tracking method of the lot as it involves calculating the correlation between all the independent variables and provides for weight updation to ensure that there is never divergance by a large amount from the actual value. The model however is lacking when it comes to the consistent tracking of peaks. The value of the mean squared error here is $J(\theta) = 0.09859203$. The next set of graphs are plotted using the concept of time series input to the model.

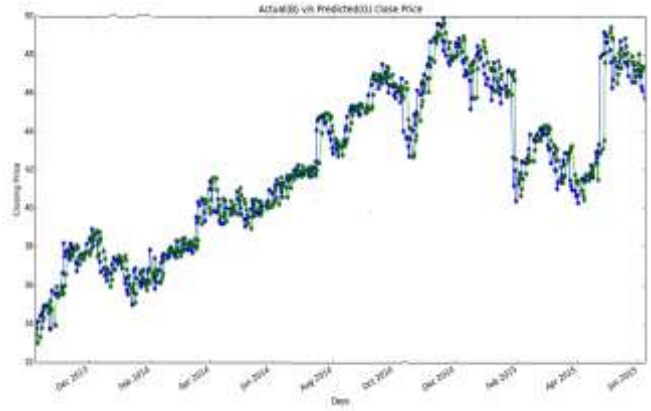

Fig 3. Linear Regression-using closing price time series to predict closing prices

The tracking is not so efficient here as it only finds the correlation between the close prices and not other independent variables. The value of the mean squared error in this case jumps to$J(\theta) = 0.41377083$

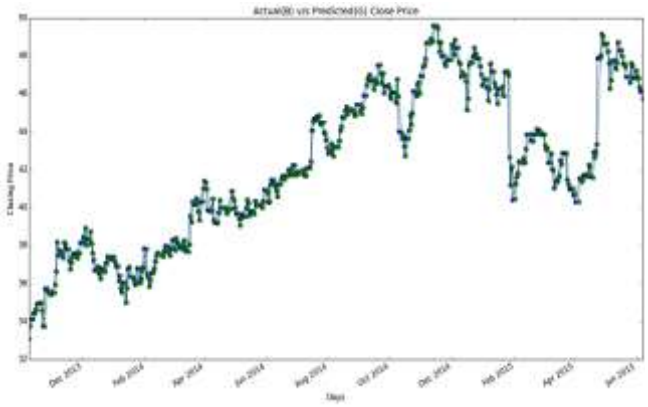

Fig 4. Linear Regression-using closing price time series to predict opening prices

The tracking in this case is much better because there is not much change in the market taking place between the period that the market closes on a day and reopens the next morning. The mean squared error $J(\theta) = 0.13984455$

5.2 Prony's Algorithm Results

We begin with using multiple independent variables to predict the dependent variable. Hence as in Linear Regression, it involves using previous day's opening price, high price, low price and adjusted closing price as independent variable to predict the dependent variable i.e. the current day's closing price.

Fig 5. Prony's algorithm-using multiple independent variables

This method converges to the optimum value of the cost function in a single step. The values of the means squared error is $J(\theta) = 0.37789569$. The primary realization is that this is far higher than what was obtained with Linear Regression with the same data set. This suggests diminished absolute accuracy in of Prony's algorithm in comparison with Linear Regression.Further comparing with Linear Regression, it is evident that there is a far more amplified lag in the predicted output. There is no way for the weights to update themselves in Normal Equation method and this might accentuate the divergence.

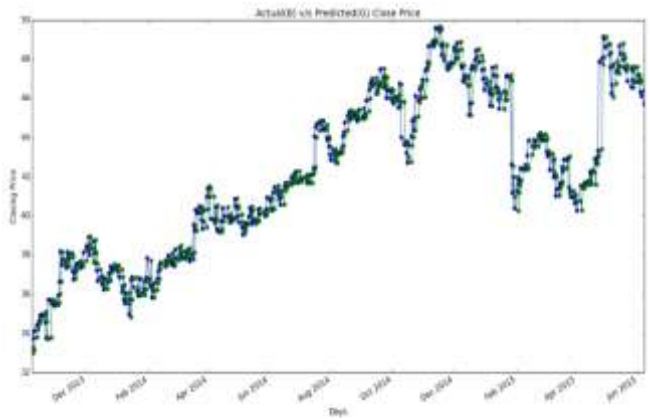

Fig 6.Prony's algorithm-using closing price time series to predict closing prices

The tracking in this case, compared to Linear Regression under the same circumstances, is far more efficient which suggests that Prony's Normal Equation method is more accurate when working with time series data. The cost function $J(\theta) = 0.3780744$

Fig 7.Linear Regression-using closing price time series to predict opening prices

The value of the cost function is $J(\theta) = 0.06922734$. This reinforces the aforementioned claim that the analysis of time series input using Normal Equation method is more accurate. It also lays credence to our earlier assertion i.e. the superior short term prediction ability of Prony's algorithm in comparison to Linear Regression.

6. FUTURE SCOPE

The magnitude of some of the peaks weren't perfectly traced. This is because those sporadic changes might have been caused by other parameters like, the volume of shares being bought increases, or the company declares dividend. Arrangements will be made to include such parameters into the model to provide better prediction. This would require normalization of data before it is fed to the system. The equation of the hypothesis function can take the form of complex curves, thus improving the accuracy. In addition, Artificial Neural Networks can be used to accurately predict the movement or the trend by using logistic activation functions or the exact values by using more advanced optimization techniques.Another avenue that can be explored is Autoregressive Integrated Moving Average

Model(ARIMA). The model is quite efficient in modeling non stationary data.

7. ACKNOWLEDGEMENTS

Sincere thanks to Prasad Joshi sir, Head of Department of Electronics at D.J.Sanghvi College of Engineering, for his unwavering support and encouragement. His guidance has been invaluable in the manifestation of these nebulous ideas. Heartfelt thanks and appreciation to Prof. Narasimhan Chari, whose sound counsel was sought at key junctures

8. REFERENCES

[1] L., Giorno, C., & Richardson, P. (1998), *Stock market fluctuations and consumption behavior: some recent evidence* (No. 208). OECD Publishing

[2] Mojaddady, Nabi and Khadivi, *Stock Market Prediction using Twin Gaussian Process Regression Modeling*, Department of Computer Engineering,Amirkabir University of Technology, Tehran, Iran

[3] Hayes, M. (2002), *Statistical Digital Signal Processing and Modeling*, Pg 144, Singapore: John Wiley & Sons (Asia) Pte. Ltd.

[4] Hayes, M. (2002), *Statistical Digital Signal Processing and Modeling*,Pg 145, Singapore: John Wiley & Sons (Asia) Pte. Ltd.

[5] Dutta, Bandopadhyay, and Sengupta, *Prediction of Stock Performance in the Indian Stock Market Using Logistic Regression*, International Journal of Business and Information, Volume 7, Number 1, June 2012

[6] Tiong, Ngo, Lee, *Forex Trading Prediction using Linear Regression Line, Artificial Neural Network and Dynamic Time Warping Algorithms*, 4 th International Conference on Computing and Informatics, ICOCI, August 2013

[7] Gharehchopogh, F. S., &Khalifehlou, Z. A. (2012), *A New Approach in Software Cost Estimation Using Regression Based Classifier*, AWERProcedia Information Technology and Computer Science, Vol 2, pp. 252-256.

[8] Draper, N. R., Smith, H., &Pownell, E. (1966),*Applied regression analysis*, Vol 3, New York: Wiley

[9] Gharehchopogh, Bonab, Khaze, *A Linear Regression Approach To Prediction Of Stock Market Trading Volume: A Case Study*, International Journal of Managing Value and Supply Chains (IJMVSC) Vol.4, No. 3, September 2013

[10] Rencher, Alvin C.; Christensen, William F. (2012), Chapter 10, Multivariate regression – Section 10.1, Introduction, *Methods of Multivariate Analysis*, Wiley Series in Probability and Statistics 709 (3rd ed.), John Wiley & Sons, p. 19,ISBN 9781118391679.

[11] Hayes, M. (2002). *Statistical Digital Signal Processing and Modeling*,Pg 499, Singapore: John Wiley & Sons (Asia) Pte. Ltd.

[12] To establish notation for future use, we'll use x (i) to denote the "input" variables Andrew Ng, CS229 Lecture notes,pp 2-22

[13] C.Sun, *Stock Market Returns Predictability: Does Volatility Matter?*, QMSS Master Thesis, Columbia University, Feb 2008