Stochastic Signal Modeling Techniques for Stock Market Prediction

Shashank Iyer
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

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

ABSTRACT
Autocorrelation measures the degree to which a current variable is correlated to the past values. Autocorrelation can be measured by running a regression equation. The study employs this temporal correlation that exists between the various stock markets related variables to predict future trends and prices, using two stochastic signal modeling processes. Data for stocks listed on the NASDAQ was scraped from the Yahoo! Finance website. Autoregressive (AR) and Autoregressive Moving Average (ARMA) techniques have been used to predict the next day's closing price using a time series input of the previous *L* days. Autoregression models the dependence of the variable to be predicted with its own lagged terms while Autoregressive Moving Average builds on Autoregression by allowing for the introduction of the Moving Average model which includes lagged terms on the residuals. The mean square error of the two was compared. The study concludes that the two models should be used in consonance for accurately modeling the magnitude and the direction of the movement in the variable to be predicted.

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

Keywords
Stock Market Prediction, DSP, Statistical Signal Processing, Regression models, Autoregressive, Autoregressive Moving Average

1. INTRODUCTION
Financial markets such as stock market are generating constantly great volume of information needed to analysis and to produce any predicting pattern in any time. Therefore they are interesting case of using different scientific methods to development and improvement in generating techniques.[1] As market participants act on this information flow, it eventually drives market prices to more efficient values.[2] Technical analysis is based on the rationale that history will repeat itself and that and the correlation between price and volume reveals market behavior. Prediction is made by exploiting implications hidden in past trading activities and by analyzing patterns and trends shown in price and volume charts. [3]The first algorithm implemented is the autoregressive model, abbreviated as AR(p). The input to this model is a time series of the closing prices and it attempts to predict the next day's close price. It finds the autocorrelation between the various elements of the time series to calculate the poles. The value of the variable 'p' denotes the number of lag terms or poles of the transfer function. The close price is the summation of the previous 'p' day's returns multiplied by the poles-$pole_1$, $pole_2$…$pole_p$.The second algorithm implemented is the autoregressive moving averages model, abbreviation being ARMA(p,q). In addition to the time series input, the moving average of the lagged terms of the erroris also applied to this model. The model uses 'p' lag terms of the variable to be predicted and 'q' lagged error terms. The error terms are the difference between the previous day's predicted close price and the actual close price. Similar to AR(p), 'p' is also referred to as the poles and 'q' as the zeros of the transfer function. So AR(p) is a special case of ARMA (p,q).Both the models can be used for prediction of long term directional trends by modeling a random process as the response of a linear shift-invariant filter to unit variance white noise.[4]

2. LITERATURE REVIEW
The 1951 thesis of Peter Whittle, *Hypothesis testing in time series analysis*, was the first to explore the general ARMA model. It was later popularized in the 1971 book *Time Series Analysis: Forecasting and Control*authoredby George E. P. Box and Gwilym Jenkins. Besides, the time series theory gives us some insights into the serial correlation effect.[5]

MartijnCremers (2002) summarizes that most of the models in previous research take the lagged returns into account, and his model is not an exception.[6]A popular adaptation of the ARMA model, the Generalized Autoregressive Conditional Heteroskedasticity (GARCH) Model was proposed by Engle (1982) and Bllerslev (1986). The GARCH process models the variances by an ARMA type process. The model often presumes that volatility is mean reverting. Hence, using GARCH (1,1), current volatility is predicted as a function of a long term mean, one lagged term of variance and one lagged term of squared return. An attractive feature of the mode is its ability to deal with excess kurtosis in the return distribution.Nonlinear GARCH models (see Hentschel, 1995, for a survey) extend the seminal contributions by Engle and Bollerslev to incorporate the asymmetric impacts of shocks or news of equal magnitude but opposite sign on the conditional variance of asset.[7]Another model evolved from the ARMA model is the Autoregressive Integrated Moving Average (ARIMA). The general transfer function model employed by the ARIMA procedure was discussed by Box and Tiao (1975). [8]The ARIMA model transforms non-stationary data into stationary data before processing it. It's quite often used to model linear time series data.

3. DATA SET

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. This excel sheet is refreshed to pull new data. The excel sheet is converted into csv file which is accessed by algorithms. 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 Since the study explores two different approaches, to create reasonable grounds for comparison, it was thought to be judicious to work with a single stock across both algorithms. For our analysis we chose the stock of Microsoft Corporation (NASDAQ:MSFT). The 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 large abrupt variations in the stock price 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. Though the period, specifically the test portion of the period, does incorporate a change in the CEO of Microsoft, this transition was along expected lines and hence was most likely priced in by the market. And finally, Microsoft being a household brand name allows for more streamlines communication of the findings of the study to a technical audience. Closing price used relates to the Microsoft stock between 9th June 2010 and 8th June 2015- a period of five years with a total of twelve hundred and fifty nine data points. The length of the period is based on sound footing. 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). These were periods when blue chip stocks faced immense downward pressure and eventually started crumbling as well. Thus this allows for the testing of the models functionality in uncertain environments.

4. ALGORITHMS

4.1 Autoregressive Model-AR (p)

4.1.1 Principles

Stochastic models are based on the LMS algorithm. The attempt here is to have the lowest possible LMS error given by:

i. $\in_{LS} = E(\hat{x}(n) - x(n))^2$

Where $x(n)$ is the time series input to the system and $\hat{x}(n)$ is the predicted output of the system for an 'n' input system. An all pole system based on stochastic modeling techniques is called an Autoregressive filter AR (p). The poles of the filter must ensure that the predicted output is very close to the actual output and is able to replicate its power spectrum as closely as possible. Based on LMS, the poles must satisfy the condition:

ii. $\frac{d\epsilon_{LMS}}{da^*_{(p)}} = 0$

where the denominator denotes the complex conjugate of the poles $a_{(p)}$.

The outer product matrix is calculated using Hebbian learning rule and is given as $x^T x$. This matrix is used to find the autocorrelation matrix, which in turn is used to calculate the poles of the transfer function using Modified Yule-Walker equations (MYWE), whose vectorised implementation is given by:

iii. $a_{(p)} = \left(R_q^T R_q\right)^{-1} R_q^T r_{q+1}$

where R_q is the autocorrelation matrix and r_{q+1} is the vector of the target values.

4.1.2 Implementation

The input to the filter is a time series of the closing price for L number of previous days. For example, closing price of day t can be assumed to be dependent primarily on the last two day's closing prices and therefore there will be two lag terms, t-1 & t-2, or two poles of the filter.

1. A time series input of closing price is considered to have a constant mean from which the instantaneous values may deviate and return to:
$$m_x(n) = m_x$$
2. The autocorrelation depends only on the difference, (k-l). The autocorrelation of the time series 'x' is denotedby:
$$r_x(k,l)$$
3. The variance of the process is finite since the variation of closing price from the mean cannot be infinite:
$$c_x(0) < \infty$$

Since the above three conditions are satisfied, the time series input of closing price is therefore called Wide Sense Stationary (WSS).The autocorrelation matrix R_q is therefore calculated as follows:

- The values of the autocorrelation sequence at lag zero is the highest and is the mean square value:
$$r_x(0) = \mathrm{E}\{|x(n)|^2\} \geq 0$$
- Since it is a real valued and causal series, it follows that:
$$r_x(k) = r_x(-k)$$
- The value of the autocorrelation sequence at any lag k>0 is the estimated value of the corresponding element in the outer-product matrix. For example:
$$r_x(k-l) = \mathrm{E}\{x(l) * x(k)\}$$

Using the above set of canonicals, autocorrelation matrix is obtained, which is a **Hermitian-Topletiz** matrix and is given by:

iv. $$R_q = \begin{bmatrix} r_x(0) & r_x(1) & r_x(p-1=2) \\ r_x(1) & r_x(0) & r_x(1) \\ r_x(2) & r_x(1) & r_x(0) \\ \vdots & \vdots & \vdots \\ r_x(L-1) & r_x(L-2) & r_x(L-p) \end{bmatrix}$$

The above matrix is an example for a three pole system.The matrix of target vectors is calculated as:

v. $$r_{q+1} = \begin{bmatrix} r_x(1) \\ r_x(2) \\ r_x(3) \\ \vdots \\ r_x(L) \end{bmatrix}$$

Now finally, these sets of matrices are multiplied using the Yule-Walker equations as specified above.

To predict the future values of closing price, the extrapolation of the Yule-Walker equation can be used which states that[9]:

vi. $$r_x(k) + \sum_{l=1}^{p} a_p(l) * r_x(k-l) = |b(0)|^2 \ldots . k \leq L$$

$|b(0)|^2$is the gain of the filter

vii. $$r_x(k) + \sum_{l=1}^{p} a_p(l) * r_x(k-l) = 0 \ldots . k > L$$[10]

4.2 Autoregressive Moving Average-ARMA (p,q)

4.2.1 Principle

The Autoregressive Moving Averages model is also based on the least squares minimization a for a time series input x(n). It is represented in block diagram form as:

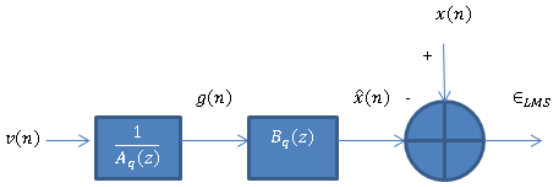


Fig1. Signal for Autoregressive Moving Average

The output of the AR process is given as input to the MA process. The MYWE for an ARMA process are given by:

viii. $$r_x(k) + \sum_{l=1}^{p} a_p(l) * r_x(k-l) = c_q(k) \ldots . k \leq L$$

ix. $$r_x(k) + \sum_{l=1}^{p} a_p(l) * r_x(k-l) = 0 \ldots . k > L$$

$c_q(k)$is the convolution of the poles with the impulse response of the filter. The signal is also a WSS as in AR.

The Topletiz matrix formation differs as follows:

R_x matrix is the collection of the autocorrelation terms from which the R_q matrix is calculated as:

x. $$R_x = \begin{bmatrix} r_x(0) & r_x(1) & r_x(p-1) \\ r_x(1) & r_x(0) & r_x(1) \\ r_x(q) & r_x(q-1) & r_x(0) \\ r_x(q+1) & r_x(q) & r_x(q-p+1) \\ \vdots & \vdots & \vdots \\ r_x(L) & r_x(L-1) & r_x(L-p) \end{bmatrix}$$

xi. $$R_q \epsilon R_x$$

xii. $$R_q = \begin{bmatrix} r_x(q) & r_x(q-1) & \cdots & r_x(q-p+1) \\ r_x(q+1) & r_x(q) & \cdots & r_x(q-p+2) \\ \vdots & \vdots & \cdots & \vdots \\ r_x(L-1) & r_x(L-2) & \cdots & r_x(L-p) \end{bmatrix}$$

The first column of R_q matrix will be selected from the second column of the R_x matrix. The number of columns is clearly dependent on the number of poles.

Using Yule-Walker equations the transfer function poles are calculated as:

xiii. $$\begin{bmatrix} r_x(q) & r_x(q-1) & \cdots & r_x(q-p+1) \\ r_x(q+1) & r_x(q) & \cdots & r_x(q-p+2) \\ \vdots & \vdots & \cdots & \vdots \\ r_x(L-1) & r_x(L-2) & \cdots & r_x(L-p) \end{bmatrix} \begin{bmatrix} a_p(1) \\ a_p(2) \\ \vdots \\ a_p(p) \end{bmatrix} = \begin{bmatrix} r_x(q+1) \\ r_x(q+2) \\ \vdots \\ r_x(L) \end{bmatrix}$$

For calculation of zeros, we need to calculate the $c_q(k)$ matrix as follows:

xiv. $$\begin{bmatrix} r_x(0) & \cdots & r_x(p-1) \\ \vdots & \ddots & \vdots \\ r_x(q) & \cdots & r_x(q-p) \end{bmatrix} \begin{bmatrix} 1 \\ \vdots \\ a_p(p) \end{bmatrix} = \begin{bmatrix} c_q(0) \\ \vdots \\ c_q(q) \end{bmatrix}$$

If suppose there is only one zero, then $r_x(k)$ matrix will have only one row. All poles lie inside the unit circle. The MA part can be found by performing a spectral factorization of the $c_q(k)$ matrix and is formed by replacing each zero that is outside the unit circle with its conjugate inside the circle [11]

Two equations are formed using the $c_q(k)$ and by taking the conjugate of the pole matrix by substituting all z^{-1} terms with z as follows:

xv. $$C_q(z) = c_q(0)z^0 + \cdots + c_q(q)z^{-q}$$

xvi. $$A_p(z) = 1 + a_p(1)z^1 + \cdots + a_p(p)z^p$$

This is basically the Laurent series expansion of the two matrices. These two equations are multiplied which yields a spectrum containing both causal and anti-causal terms. The coefficients of the anti-causal terms are replaced by the coefficients of the causal terms. This is termed $P_y(z)$. Finally, a spectral factorization is performed i.e. only the causal part of $P_y(z)$ is retained which gives the equation of the zeros expanded in Laurent series given by$B_q(z)$.

4.2.2 Implementation

The AR coefficients indicate the lag terms, i.e., the degree of dependence of the present day's closing price on each of the previous days' closing prices. The first closing price is predicted using only the poles. The zeros indicate the degree of dependency on the error between the previous day's predicted close price and its actual value. Adding of a pole to the system, pulls the root locus to the right of the $j\omega$ axis, which makes it unstable while adding a zero pulls it to the left thus ensuring stability. The ARMA models implemented here have equal number of poles and zeros. This error term is added to the next day's closing price.

5. RESULTS

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

5.1 Autoregression Results-AR (p)

The Autoregressive model using only one pole(not shown) will always have a debilitating lag in its predicted output because it implies that the today's close price is almost equal to the yesterday's close price.

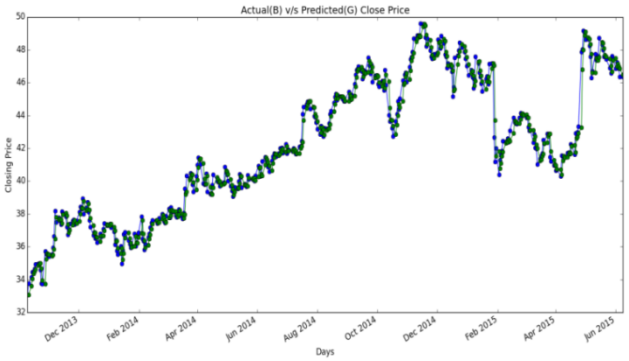


Fig 2.AR(2) model used to predict the present day's closing price

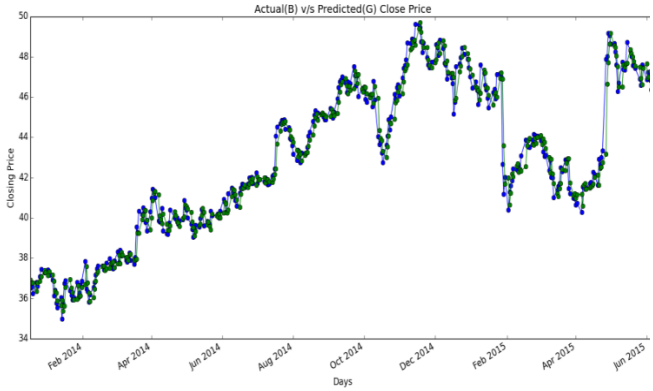


Fig3.AR (8) model used to predict the present day's closing price

On comparison it is clear that using fewer poles is more accurate than using more number of poles or lag terms. This is clearly illustrated in the graph below:

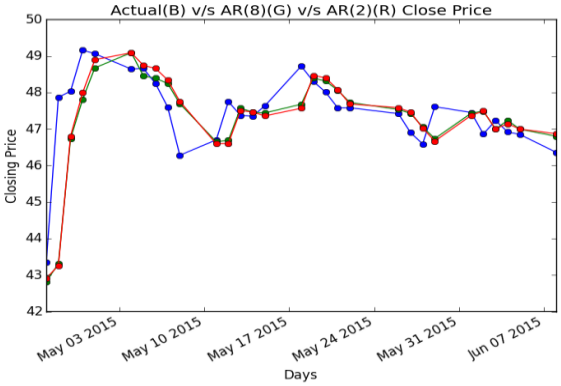
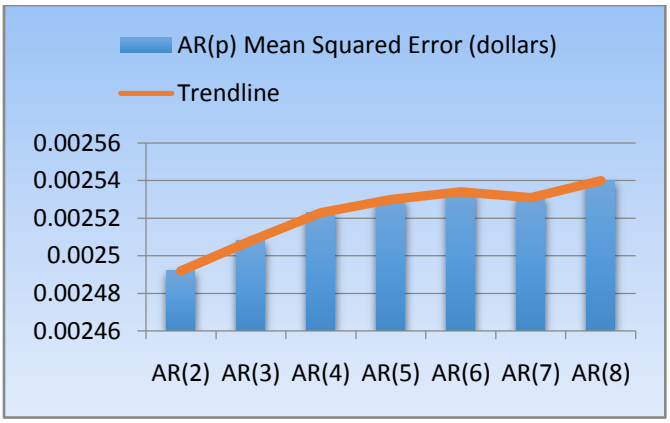


Fig 5. A magnifiedview of AR(2) (Red) and AR(8) (Green) plots superimposed over the actual closing price (Blue)

The comparison between the various AR models indicates that AR(2) has the least error and is best suited for use

Table 1.A statistical **comparison of various AR Models**



Referring to the number of lags included, there is no agreement reached, but a quite small number would be preferred to avoid data snooping problem. [12] Data snooping is a model over fitting problem first discussed by Lo and MacKinlay (1990). This often creates persistence in models or what is colloquially referred to as lag. This tendency in the model to mirror the previous day's direction of stock price movement would make impair its ability to track the peaks and troughs of stock price movement.

5.2 Autoregressive Moving Average Results-ARMA (p,q)

ARMA model alleviates the overfitting problem evident in the AR model. ARMA models can be used to remove persistence.

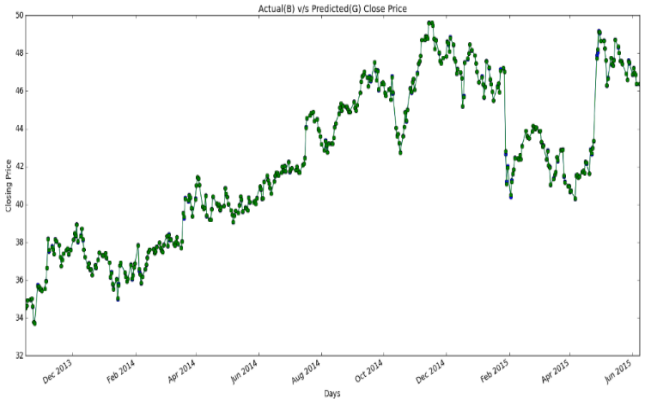


Fig6.ARMA(1,1) used to predict the present day's closing price

From the graph above, it is very clear that tracking of the peaks is very efficient with the inclusion of the moving average term.

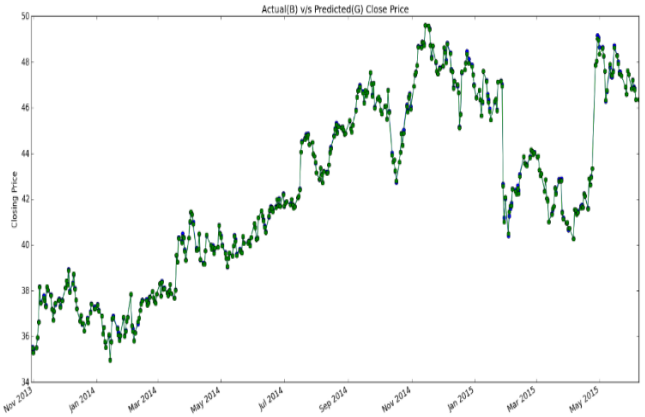


Fig 7.ARMA(8,8) used to predict the present day's closing price

If the number of lag terms is increased, it causes over fitting.Using more lag variables also results leads to data snooping and erroneous outputs. This is indicated below by observing the comparisons of lower and higher order ARMA models.

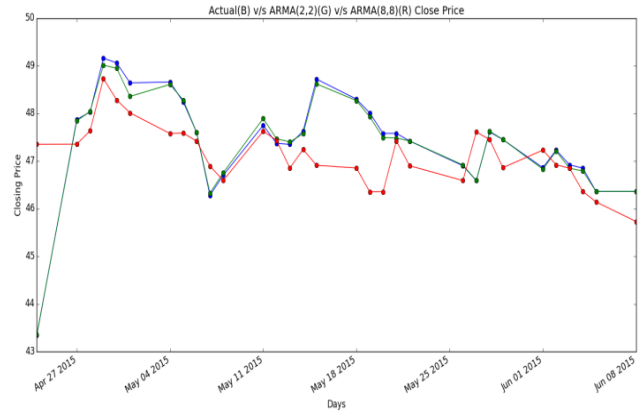


Fig 8.A magnifiedview of ARMA (2,2) (Green) and ARMA (8,8) (Red) plots superimposed over the actual closing price (Blue)

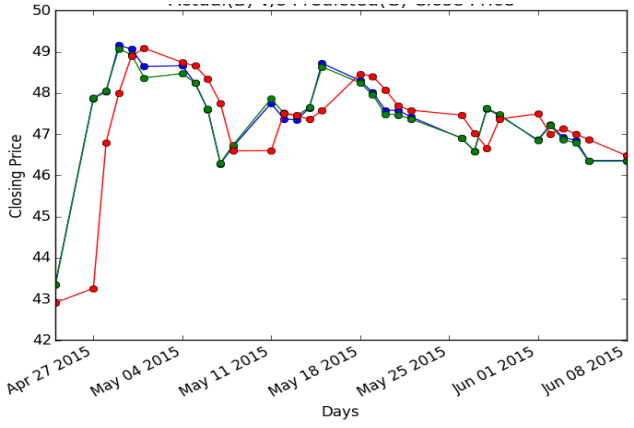
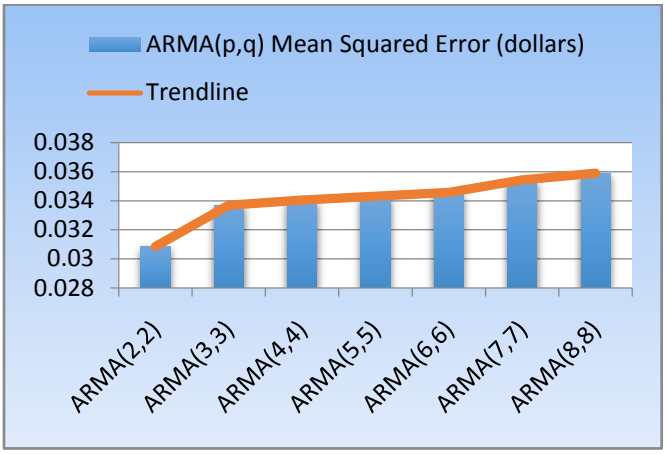


Fig 9. A magnified view of ARMA(2,2)(Green) and AR(2)(Red) plots superimposed over the actual closing price(Blue)

The plot above indicates that since ARMA uses correction terms, it is able to overcome any lag in its predicted output which AR may not be able to as effectively.

Table 2.A statistical comparison of various ARMA Models



6. FUTURE SCOPE

The mean squared error obtained for the AR and ARMA processes suggests that the AR process is more accurate over than ARMA. However through visual analysis we infer that ARMA tracks the peaks and troughs far more efficiently. Thus opening up an opportunity to explore the amalgamation of these two models that minimizes the mean squared error while retaining the ability of the ARMA model to shake off the over fitting problem. Another avenue we are excited at exploring is the use of classification algorithms such as Fisher linear discriminant analysis, K-nearest neighbor and Support vector machine in consonance with these stochastic models. The classification algorithms could be used to distribute a large tranche of data into categories. Separate regression coefficients could be calculated for each of these categories. New data could then be categorized into one of these groups and the relevant regression coefficient used for more accurate prediction. In essence, using we attempt to match data to similar historical trends.And finally, far more sophisticated methods of judging models such as the Log Likelihood Ration could be explored to offer different perspective.

7. ACKNOWLEDGMENTS

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] 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.

[2] Lochmiller, Chen, *Predicting Short Term Stock Returns*, CS299 Stanford, December 2013

[3] Vaisla, Bhatt, *An Analysis of the Performance of Artificial Neural Network Technique for Stock Market Forecasting*, International Journal on Computer Science and Engineering Vol. 02, No. 06, 2010, 2104-2109

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

[5] Campbell, J. Y., and M. Yogo, 2006, *Efficient tests of stock return predictability*, Journal of Financial Economics, 81 (2006): 27-60.

[6] MartijnCremers, K. J., 2002, *Stock Return Predictability: A Bayesian Model Selection Perspective*, The Review of Financial Studies, Vol. 15, No. 4. (Autumn, 2002), pp. 1223-1249.

[7] Arowolo,W.B, *Predicting Stock Prices Returns Using Garch Model.* The International Journal of Engineering and Sciences, Vol. 2 Issue 5 (2013) Pages 32-37, 2319 – 1813

[8] Box, George EP, and George C. Tiao, *Intervention analysis with applications to economic and environmental problems*, Journal of the American Statistical Association 70.349 (1975): 70-79.

[9] Hayes, M. (2002), *Statistical Digital Signal Processing and Modeling,* Pg 194, Singapore: John Wiley & Sons (Asia) Pte. Ltd.

[10] Durbin, J. 1960,*Estimation of parameters in time-series regression models*, Journal of the Royal Statistical Society B 22

[11] Hayes, M. (2002), *Statistical Digital Signal Processing and Modeling,* Pg 195, Singapore: John Wiley & Sons (Asia) Pte. Ltd

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