Prediction Based Portfolio Optimization Model Using Neural Networks with an Emphasis on Leading Stocks of NSE

Gajendra K. Vishwakarma* and Chinmoy Paul
Department of Applied Mathematics, Indian Institute of Technology (ISM) Dhanbad-826004, India
Email: vishwagk@rediffmail.com, chinmoy.gcc@gmail.com

Abstract

This work discusses most frequently traded stocks of National stock exchange of India. A prediction based portfolio optimization model is considered to present an ideal portfolio out of the considered stocks. Neural network has been used to predict stock returns and a risk measure is derived that has the same foundation as that of mean variance model. The architecture of the network is designed experimentally. The cross-correlation structure between the fluctuations of price for frequently traded stocks in national stock exchange has been studied for a period of last one year. The structure of interactions is studied using the spectral properties of the cross-correlation matrix. Spectral decomposition of the cross correlation matrix is used in finding the most influencing underlying stocks of NIFTY 50. The portfolio formed by the presented model is considered to verify the presence of influencing stocks as determined by the spectral analysis.

Keywords: Portfolio Selection; Time series prediction; Neural Network; Spectral Decomposition.

1. Introduction

National Stock Exchange (NSE) and Bombay Stock Exchange (BSE) are the leading stock exchange in India. These two exchanges have brought about unparalleled transparency, speed and efficiency, safety and market integrity. Now with the growing digitalization and ease of banking facilities customer is slowly and gradually getting attraction towards online trading. Online trading accounts are the platform where a bank provides its customer the facility to invest in the share market without having the trouble to find a broker. Before making an investment an individual will like to optimize his portfolio in terms of minimizing risk and maximizing return. Selection of instrument for investment is a crucial task in financial theory and practice and it is basically aligned with the future performance of the selected instruments, mainly their expected returns. When investments are exposed to uncertainties, the investment selection framework must include a quantitative measure of the uncertainty of obtaining the expected return, i.e. a quantitative measure of risk.

The return of an index depends on the performance of the underlying constituents. A landmark in modern portfolio theory is the Mean Value model proposed by Markowitz (1952, 1991). In mean variance model, risk in an investment is minimized by optimal selection of instruments with low joint risk, which provides a mechanism of loss compensation known as efficient diversification. Portfolio optimization procedure consists in selecting a number of stocks, the participation of each stock that minimizes risk with a given level of return. Many other models are developed Elton et. al. (2009) using the fundamental assumption of Mean variance model. The assumption of the Mean Variance model is often violated by many real word examples. One such violation is deviation from normality by the returns of the stock Fama (1965).

The predictability of stock markets is still an open area in finance theory. The efficient market hypothesis (EMH), the theoretical framework that guides the discussion around this question, has been under empirical testing and reviewing for the past several decades Malkiel and Fama (1970), Fama (1991, 1998), Malkiel (2003). Market efficiency implies a random walk model for the prices of stocks, but pricing irregularities and predictable patterns like serial correlations, calendar effects, and even sports results effects do appear Malkiel (2003), Edmans et.al. (2007).

Forecasting in time series was customarily attempted by the linear models of time series analysis Box et. al. (2015). However, if the nature of time series is nonlinear prediction using linear models may lead to results far away from the expected. In recent past many machine learning approach such as neural network, support vector machine, fuzzy sets (White (1988), Huarng et. al. (2007)) for time series
forecasting has come up. Among these approaches neural network approach is advantageous for forecasting and estimation of model parameters due its ability to estimate the functional relationship between input and output variables be it linear or non-linear.

Fernandez and Gomez in (2007) presented a heuristic method based on artificial neural network to focus on the problem of tracing out the efficient frontier for the general mean variance model with cardinality and bounding constraints. Hopfield network had been used to the portfolio selection problem. All the experiments showed that the neural network model has given better solutions than the other three heuristic methods. Thawornwong and Enke (2004) presented a data mining approach to choose economic and financial variables and used the selected variables with predictive power as input of a neural predictor. The neural predictor is used in predicting the direction (sign) of future return movements. The predictions were used to implement a trading strategy for deciding either to invest on the SandP500 index portfolio or on the T-Bill (risk-free) during one month. The result of their study showed that the monthly returns obtained with this adaptive strategy were always higher than those of the (non-adaptive) compared methods.

Niaki and Hoseinzade (2013) used artificial neural network to forecast the daily price change direction of standard and poor’s 500 (SandP 500) index. Factorial design of experiment considering 27 potential financial and economical variables as factors was applied to find the most influential features for the proposed ANN. The results of their proposed approach showed that the ANN that uses the most influential features is able to forecast the daily direction of SandP 500 significantly better than the traditional model. Sinha and Kovur (2014) in their work analyzed cross correlation between the fluctuations of 74 currencies. The Eigen value analysis of the cross correlation matrix presented that bulk of the Eigen values falls within the bounds predicted by random matrix. However, the ones deviating from the bulk over the bounds contain the important information about groups of strongly interacting components. This idea is used in finding the most influential stock among the underlying 50 stocks of the index Nifty-50 and is compared with the portfolio formed by portfolio selection model suggested by Freitas et. al (2009).

In this work, we present an approach to select optimal portfolio out of a given set of instruments. The portfolio optimization is based on the predictions using artificial neural network. Another approach based on the Eigen values of the cross correlation matrix is also presented. Both the method suggests an efficient portfolio out of a given set of instruments. A comparison between both the methods has also been carried out.

2. Cross correlation matrix for the return

To quantify the degree of association-ship between the underlying of Nifty 50 index, first the return for each underlying instrument is calculated

\[
R_i(t, \Delta t) = P_i(t + \Delta t) - P_i(t)
\]

where \(P_i(t)\) is the price \(i^{th}\) underlying security at time \(t\). Prices considered here are the daily closing price. \(\Delta t\) is considered as one day. The standard deviation of the returns and normalized returns are obtained as \(\sigma_i = \sqrt{(R_i - \bar{R})^2}\) and \(r_i(t, \Delta t) = R_i / \sigma_i\), respectively.

After obtaining the normalized returns for all \(N\) underlying over a period of \(T\) days the cross-correlation matrix \((C)\) between the normalized returns is computed as \(C_{ij} = \langle r_i(t), r_j(t) \rangle\). If the movement of the different underlying instruments is uncorrelated, the resulting random correlation matrix (referred to as a wishart matrix) has eigen values distributed according to Sengupta and Mitra (1999).

\[
D(\lambda) = \frac{Q}{2\pi} \sqrt{(\lambda_{\text{max}} - \lambda)(\lambda_{\text{min}} - \lambda)}
\]

As \(N \to \infty\) and \(T \to \infty\) with \(Q = (T/N) > 1\) the distribution of the eigen values of the normalized returns are given by \(\lambda_{\text{max}} = [1 + (1/Q)]^2\) and \(\lambda_{\text{min}} = [1 - (1/Q)]^2\). The data that has been analyzed the value of \(Q\) is 4.92 and \((\lambda_{\text{max}}, \lambda_{\text{min}}) = (1.45, 0.635)\). The distribution of the eigen value is shown in Fig.1.
3. Prediction Based on Optimization of Portfolio

Mean Variance theorem by Harry Markowitz has received considerable research attention Markowitz (1956), Hamza and Janssen (1996), Sharpe (1963), Konno and Yamazaki (1991), Sortino and Van Der Meer (1991) due to its computational feasibility, model simplifications and the development of risk measures. The prediction of future return in the context of portfolio selection has received little attention. In most of the models same prediction method is employed as in the case of Mean Variance theorem i.e. the mean of the past return. However, mean returns are expected to be verified only in case of long term predictions and show inadequacy in short term prediction of future return. Here an alternative procedure has been presented that describes prediction based portfolio optimization model using predicted returns as the expected return. Unlike mean variance theorem, this procedure uses the variance of the errors of prediction as risk measure.

3.1 Expected return and risk of a stock.

Let the return of a stock and its expected returns are denoted by $r_i$ and $\tilde{r}_i$ the relation between them is given by

$$r_i = \tilde{r}_i + \epsilon_i,$$

where $r_i$ is the stock return at time $t$, $\tilde{r}_i$ is the predicted return for time $t$ obtained at $t-1$ and is $\epsilon_i$ the error at time $t$. Rearranging the terms, the error at time $t$ can be presented as

$$\epsilon_i = r_i - \tilde{r}_i.$$  

For an unbiased predictor the series of error $\epsilon' = (\epsilon_1, \epsilon_2, ..., \epsilon_n)$ prediction must be statistically independently and identically distributed with mean and variance given by

$$\mu = E(\epsilon_i) = 0$$  

and

$$\nu = \sigma^2 = \frac{1}{n-1} \sum_{i=1}^{n} \epsilon_i^2.$$  

The variance equation gives the amount of uncertainty about the realization of predicted returns. Higher the variance higher is the risk. The variance component is used in modeling the optimization models for a portfolio.

3.2. Expected return and risk of a portfolio

A portfolio is a collection of $P$ stocks and the corresponding weights $(\omega_i)$ in such a way that each of the weight $\omega_i$, $i = 1, 2, ..., P$, $0 \leq \omega_i \leq 1$ and $\sum_{i=1}^{P} \omega_i = 1$ represents fraction of portfolio invested in stock $i$. The expected return of the portfolio or the predicted return of the portfolio $\tilde{r}_p$ is the linear combination of expected return of each of the stocks and the corresponding weights.

$$\tilde{r}_p = \sum_{i=1}^{P} \omega_i r_i$$
With the assumption that return of each stock follows normal distribution, portfolio risk is considered as the variance of the joint distribution of the linear combination of expected return of each stock and its corresponding weights.

\[ V = \text{Var} \left( \sum_{i=1}^{p} \omega_i r_i \right) = \sum_{i=1}^{p} \sum_{j=1}^{p} \omega_i \omega_j \psi_{ij} \]  

(8)

Where \( V \) is the total portfolio risk, \( P \) is the number of stocks and \( \omega_i, \omega_j \) are participation of the stock \( i \) and stock \( j \) respectively. \( \psi_{ij} \) represents the covariance between the expected returns of the stock \( i \) and stock \( j \) respectively and is given by

\[ \psi_{ij} = \psi_{ji} = \frac{1}{n-1} \sum_{t=1}^{n} \varepsilon_i \varepsilon_j \]  

(9)

Equation (9) can also be presented as

\[ V = \sigma_p^2 = \sum_{i=1}^{p} \omega_i^2 \sigma_{ei}^2 + \sum_{i=1}^{p} \sum_{j=1}^{p} \omega_i \omega_j \psi_{ij} \]  

(10)

here the first term represents risk associated to each of the stocks to the portfolio and the second term represents interactive predictive risk of the presence of each pair of the stock \( i \) and \( j \).

3.3 Optimization of the portfolio

Keeping all the assumptions intact the optimization model can be presented as

\[ \text{Minimize} \quad V = \sum_{i=1}^{p} \omega_i^2 \sigma_{ei}^2 + \sum_{i=1}^{p} \sum_{j=1}^{p} \omega_i \omega_j \psi_{ij} \]  

(11)

Subject to

\[ \sum_{i=1}^{p} \omega_i r_i = R_d \]  

(12)

\[ \sum_{i=1}^{p} \omega_i = 1 \quad ; \quad 0 \leq \omega_i, \quad i = 1, 2, ..., p \]  

(13)

The formulated optimization problem minimizes the risk (Eq. 11) at a desired return (Eq. 12) and Eq. (13) represents the restriction of the model to purchase only.

4. Experiment

This section presents the experiment that has been carried out using the underlying constituents of Nifty-50 index. The method used to predict the returns along with the measures to verify the performance of the prediction based portfolio optimization model are also presented.

4.1 Data

We have considered daily closing prices of 50 companies listed under Nifty-50. This data is commonly available in the websites of BSE or NSE (https://www.nseindia.com, http://www.bseindia.com). A period of 1 year span has been considered starting from 17 December 2015 till 15 December 2016. A crucial point in collecting data for the constituents of Nifty-50 index is that, the index is reviewed after every six months and based on the decision of index policy committee the underlying constituent changes. Here we have considered the same set of index constituents that were present at during the period of July-December 2016. Historical data for the same set of constituents has been considered.

4.2 Prediction of returns

Backpropagation network has been used for predicting the returns Wong (1991), Zhang et.al. (1998), Ghiassi and Saidane (2005). The architecture of the network is designed experimentally. Depending upon the complexity of the problem suitable network architecture is designed. For checking the normality of the error of the predicted returns the last step of our proposed algorithm (Paul and Vishwakarma 2016) has been used. This imposed condition not only checks normality of the error of
the predicted returns but also helps in finding suitable network architecture for a particular input data set.

Entire data set is divided in three parts, 50% of the data is used in training, 30% validation and 20% for testing. The use of training and validation is to control overfitting. Haykin (1999) presented incompetency in controlling overfitting for non linear time series prediction due to its structural regime change. Though there is no hard and fast rule available for controlling overfitting a better suitable technique like shifting the validation segment to other locations in time Pantazopoulos et. al. (1998) is used. The overfitting controlling procedure we have used is inspired in the Non-linear Cross Validation (NCV) proposed in Moody, J. (1994). The training and testing procedure described above is repeated for all 50 underlying of the Nifty-50 index by advancing a sliding widow of twenty days one day at a time.

4.2.1 Evaluation Matrics

The performance evaluation measures that are used in this study are Mean Error (ME), Root Mean Square Error (RMSE), Absolute percentage Error (APE) and Hit Rates Matrics (HRM).

Mean Error is the average of the difference of actual and the predicted returns and is given by

\[
ME = \frac{1}{n} \sum_{t=1}^{n} (r_t - \tilde{r}_t)
\]

(14)

\(r_t\) and \(\tilde{r}_t\) are the realized and expected returns of a time series at time \(t\) and \(n\) is the length of the time series. ME is basically used to verify the assumption that the error of the expected return follow normal distribution with zero mean and a fixed standard distribution.

RMSE is used in comparing two time series in terms of its variability. It is well known that RMSE is badly affected by the presence of outliers, hence a time series with outlier will have more RMSE compared to other and is defined as

\[
RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^{n} (r_t - \tilde{r}_t)^2}
\]

(15)

The APE is defined as

\[
APE = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{r_t - \tilde{r}_t}{r_t} \right|
\]

(16)

APE is unit free and averages out very large values and very sensitive to very small changes in data. Hit rate \(H_R\), \(H_{R^+}\), \(H_{R^-}\) measures the rate of coincidence of actual return \(r_t\) and \(\tilde{r}_t\) the predicted return. \(H_R\) measures the percentage where both \(r_t\) and \(\tilde{r}_t\) have same signal and are different from zero. \(H_{R^+}\) measures the percentage where both are positive and \(H_{R^-}\) measures the percentage where both have same signal and are negative.

\[
H_R = \frac{1}{n} \sum_{t=1}^{n} I_t, \quad H_{R^+} = \frac{1}{n} \sum_{t=1}^{n} I_t, \quad H_{R^-} = \frac{1}{n} \sum_{t=1}^{n} I_t
\]

The heat rate measure will give us the idea of the underlying stocks that that are with positive return, normal return and negative return. So, the search for the effective underlying can be traced out with the help of hit rate measures. The underlying time series with maximum hit rate should be included in the portfolio to achieve maximum return.

5. Results and Discussion

Portfolio selection is an open problem and there are enormous research articles suggesting methods to select portfolio. The main target here is to find a set of underlying from a set of available instruments. The set of available instruments here are 50 underlying of Nifty-50 index. Various
evaluation measures discussed in previous section i.e. ME, RMSE, APE, \( H_R, H_K, H_K' \) are presented in Table-1 (appendix). Value of \( H_R \) above 50% indicates that the predictor achieved a performance above the pure chance of predicting the positive return of the market. Instruments with high percentages of \( H_K' \) are highlighted in color. From the distribution of eigen value it is observed that bulk of the eigen values are beyond the defined limit and Maximum eigen value is around 10 times the upper bound. Eigen vectors corresponding to first four (magnitude wise) eigen value are presented in Fig-2 and forecasting return time series using neural network shown in Fig-3. If we closely consider the distribution of eigen vector corresponding to 1\(^{st} \) eigen value, it may generally be stated that higher weights are given to those instruments which are having higher percentages of \( H_K' \).

![Fig 2: Eigen Vectors Corresponding to first four eigen value](image)

![Fig 3: Forecasting return time Series using Neural Network](image)

5. Conclusion:

In this article underlying instruments of Nifty-50 index are analyzed to form a portfolio that can maximize return with minimum risk. Two different approaches one based on cross correlation structure of the return matrix and the other based on prediction using neural network are discussed. Maximum eigen value of the cross correlation matrix explains the maximum variability of the return data. The corresponding eigen vector gives higher weight to those underlying which are expected to contribute towards portfolio formation. Underlying instruments having higher weight and the instruments having high percentages of are \( H_K' \) almost same. Hence based on the present study any one of the approach can be adopted for a portfolio selection in case of predefined set of underlying are available at hand The set of instrument considered in this study are from a well-known index. The return of the underlying time may depend on other potential influential features which are beyond the scope of this study. Only partial results based on \( H_K' \) and cross correlation matrix are presented which provide an initial structure of a portfolio with expected positive return, the final step i.e. minimization of risk is under exploration.
References