Golden Research Thoughts
ISSN 2231-5063
Volume-3 | Issue-10 | April-2014 Available online at www.aygrt.isrj.net

PLS COMPUTATIONAL MODEL FOR PREDICTION IN TIME SERIES

S. Meenakshi Sundaram and M. Lakshmi

Research Scholar, Faculty of Computing Sathyabama University Tamilnadu, India.
Head, Faculty of Computing , Sathyabama University , Tamilnadu, India.

Abstract : Some data fluctuates rapidly in a short period of time. Classical and computational models are useful in predicting this highly volatile data. In this study Partial Least Square Regression and computational neural network models are used to explore stock market tendency. Thirteen variables are considered to predict the daily closing prices of BSE sensex data. To evaluate the prediction ability of the models, standard error values are calculated. The results revealed that Nonparametric PLS regression model is better in prediction.

Keywords: PLS Regression, Multivariate Normality, Root Mean Square Error, Prediction, MLP.

INTRODUCTION

Stock Index prediction has gained importance in recent times due to its commercial application and the nature of the data explored. The noisy environment of the data attracts many researchers to explore it. Many technical indicators are available to predict the daily prices. Each index has its own meaning and interpretation. In this study the indicators considered were daily opening price, high price, low price, volume of transaction, Adjusted Closing prices, Stochastic Oscillator, Rate of convergence, Moving averages, disparity and momentum for predicting the closing prices from 1st January 2008 till 31st December 2010(737).

All traditional parametric time series models like ARIMA, GARCH are valid only when the normality assumption is itself justified. Testing of normality have been proposed Mardia [6],[7] and D'Agostino [16]. The need for testing normality in a multivariate setting is discussed by Ganadesikan (1977), Cox and Small [1] and Cox and Wermuth [2]. Multivariate Omnibus test proposed by Mardia was considered to check the condition of normality of the predictors.

Violation of normality results in using the Nonparametric Models as the only alternative. Partial Least Square regression, Principal Component Analysis, Ridge Regression models are some of the Classical Statistical Models and Neural Network is the most widely used computational non parametric model. Artificial Neural Network has been widely applied in forecasting financial time series data without any parametric assumptions. Myungsook [15] has investigated the effectiveness of inputs in market prediction. Other techniques involve Hybrid Kohonen Self Organisation Map (SOM)[14], Hybrid Financial Systems [17], GA based Support Vector Machines[12], TSK fuzzy rule based systems[16]

MULTIVARIATE NORMALITY

Multivariate normality plays a crucial role in analyzing data using Statistical procedures. Many Multivariate normality tests are available to test the Marginal normality in a multivariate data. The Shapiro-Wilk test is used if sample size is less than or equal to 5000, otherwise, the Lilliefors test (Kolmogorov-Smirnov test with estimated parameters) is favored. Mardia's skewness and kurtosis coefficients are computed and tests of significance are performed for these coefficients using asymptotic distributions. These tests are generally effective for testing multivariate normality. Henze-Zirkler test statistic is found with the associated p-value using the lognormal distribution. Finally, the beta Q-Q plot of scaled squared Mahalanobis distances is plotted following the approach of Gnanadesikan and Kettenring (1972).

The test for multivariate normality proposed by Mardia [6][7] which was used to measure the multivariate skewness and kurtosis and joint Multivariate normality. The Henze-Zirkler test is based on a nonnegative functional D(.,.) that measures the distance between two distribution functions and has the property that $D(N_d(0, I_d), Q) = 0$ if and only if $Q = N_d(0, I_d)$ where $N_d(\mu.\Sigma\, d)$ is a *d*-dimensional normal distribution. The distance measure *D*(.,.) can be written as

$$D_\beta(P, Q) = \int_{R^d} \left|\hat{P}(t) - \hat{Q}(t)\right|^2 \varphi_\beta(t$$

where $\hat{P}(t)$ and $\hat{Q}(t)$are the Fourier transforms of P and Q, and$\varphi_\beta(t)$ is a weight or a kernel function. The density of the normal distribution $N_d(0, \beta^2 I_d)$is used as

$$\varphi_\beta(t) = (2\pi\beta^2)^{\frac{-d}{2}} \exp\left(\frac{-|t|^2}{2\beta^2}\right), t \in R^d$$

where $|t| = (t' t)^{0.5}$. The parameter βdepends on *n* as

$$\beta_n(n) = \frac{1}{\sqrt{2}}\left(\frac{2d+1}{4}\right)^{\frac{1}{(d+4)}} n^{\frac{1}{(d+4)}}$$

The test statistic computed is called $T_\beta($ and is approximately distributed as a log normal. The log normal distribution is used to compute the null hypothesis probability.

PRINCIPAL COMPONENT ANALYSIS

Principal component analysis (PCA) technique consists in rewriting the coordinates in a data set in other coordinates system which will be more convenient for analysis. This new coordinates are represented on orthogonal axis, being obtained in decreasing variance order. The total amount of principal components is equal to the amount of original variables and presents the same statistical information. The PCA is defined as follows:

Let $X' = (x_1, x_2, \dots, x$ be a p dimensional random variable. The principal component of is
$y_j = e_j' x = e_{1j} x_1 + e_{2j} x_2 + \cdots + e_{pj} x_p\ , i = 1,2, \dots, p, e_j'\, e_j = 1$

and it must satisfy the following conditions: The variable is the one whose variance is maximum among all of the variance of $y =$. The variable is not correlative with $y_1, y_2, \dots, y_k\ (k = 1,2,3, \dots, p -$

PARTIAL LEAST SQUARE REGRESSION

Partial Least Square Regression generalizes and combines the features of principal component analysis and multiple regressions which is obviously related to canonical correlation and to multiple factor analysis. The main originality of PLS regression is to preserve the asymmetry of the relationship between predictors and dependent variables, whereas these other techniques treat them symmetrically. In partial least squares regression, prediction functions are represented by factors extracted from the ***Y'XX'Y*** matrix. The number of such prediction functions that can be extracted typically will exceed the maximum of the number of *Y* and *X* variables.

Furthermore, partial least squares regression can be used as an exploratory analysis tool to select suitable predictor variables and to identify outliers before classical linear regression. In PLS the factors to predict the responses in the population are achieved indirectly by extracting latent variables and from sampled factors and responses respectively. The extracted factors are used to predict the Y scores and then the predicted Y-scores are used to construct predictions for the responses. In PLS the X- and Y-scores are chosen so that the relationship between successive pairs of scores is as strong as possible. In principle, this is like a robust form of redundancy analysis, seeking Directions in the factor space that are associated with high variation in the responses but biasing them toward directions that are accurately predicted.

MULTILAYER PERCEPTRON

Neural Networks[NN] have been used in function estimation such as stock price prediction, option price modeling, portfolio optimization and currency exchange rate estimation (Steiner and Wittkemper [13]; Yao and Tan [5]; Galindo [3]; Leigh et al. [20]; Hutchinson et al. [4]; Trafalis et al. [9]. NN is a learning machine that is designed to model the way in which the brain performs the particular tasks. The multi-layer perceptron (MLP) is the most widely used type of NN for function approximation. Input quantities are processed through successive layers of "neurons". Each neuron of a layer other than the input layer computes first a linear combination of the outputs of the neurons of the previous layer, plus a bias. The coefficients of the linear combinations plus the biases are called the weights. Neurons in the hidden layer then compute a non-linear function of their input. The two main activation functions used in current applications are hyperbolic tangent and sigmoid in which hyperbolic tangent ranges from -1 to 1, and the latter is equivalent in shape but ranges from 0 to 1.

Multilayer Layer Perceptron has rescaling option which is done to improve the network training. There are three rescaling options: standardization, normalization, and adjusted normalization. All rescaling is performed based on the training data, even if a testing or holdout sample is defined. The units in the output layer can use any one of the following activation function - Identity, Sigmoid, Softmax or Hyperbolic Tangent. Sum of square error and the relative error measures are used to find the best neural network model. Standardized Error Measures used for comparing the ability of the models are

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|A_t - P_t|}{A_t}$$

where A_t is the actual value and P_t is the predicted value.

$$RMSE = \sqrt{\frac{1}{n}\sum(A_t - P_t)^2}$$

RESULTS AND FINDINGS

To predict the daily closing prices the variables daily opening, high price, low price, volume of transaction, adjacent closing, stochastic oscillator, Rate of Convergence, momentum, moving averages, disparity and price oscillator are considered.

MULTIVARIATE NORMALITY

The marginal normality of the variables is tested using Shapiro Wilks test. Form the Table 1, the p-value of all the variables are less than 0.01, which implies that the variables are not normally distributed. The joint normality is checked by calculating Mardia's multivariate skewness, multivariate kurtosis and Henze-Zirker test values. In all three evaluations (Table 2), p-value is less than 0.01hence confirming that the multivariate normality is not met. Figure 1 gives the Q-Q plot .Since the assumption of normality is not met nonparametric models can only be used to predict the closing prices.

TABLE 1 TESTING MARGINAL NORMALITY

Variables	Test Statistic	p-value
Open	0.921	0.000
High	0.920	0.000
Low	0.922	0.000
Volume	0.699	0.000

Adj_Close	0.920	0.000
SO	0.895	0.000
ROC	0.945	0.000
Momentum	0.960	0.000
MA5	0.917	0.000
MA7	0.916	0.000
MA10	0.914	0.000
Disparty	0.942	0.000
Price oscillator	0.941	0.000

TABLE 2 MARDIA AND HENZE-ZIRKER TEST

Test	Coefficients	Test Statistic	p-value
Mardia Skewness	164.399	19931.441	0.000
Mardia Kurtosis	568.663	254.558	0.000
Henze-Zirkler		5.567	0.000

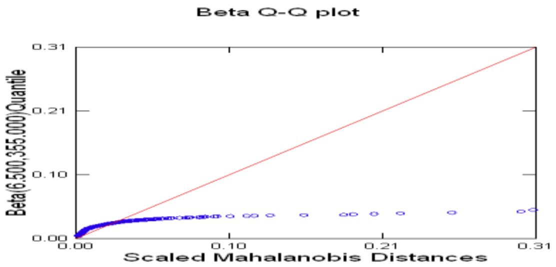


Figure 1 BETA Q-Q PLOT

PRINCIPAL COMPONENT ANALYSIS

The predictors are found to be highly collinear in nature. This could be overcome by using the multivariate techniques like Principal component Analysis, Ridge Regression, Partial least square regression and so on. To determine the number of components to be extracted from the variables taken, Scree plot is plotted. From Fig 2, two principal components accounts for maximum variability of the variables. Thus two principal components are extracted from it. Table 3 displays the coefficients of the variables.

TABLE 3 COEFFICIENT OF VARIABLES

Variables	PCA1	PCA2
Open	0.723	0.689
High	0.722	0.691
Low	0.731	0.681
Volume	-0.993	0.122

Adj_Closing	-0.728	0.684
Stochastic Oscillator	-0.019	-0.041
Rate of Convergence	-0.032	-0.014
Momentum	0.032	0.063
MA5	0.027	0.685
MA7	0.728	0.685
MA10	0.727	0.684
Disparty	0.49	0.024
Price Oscillator	0.075	0.085

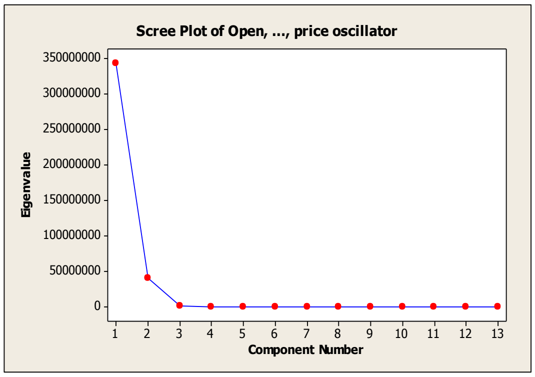


Figure 2 SCREE PLOT

PARTIAL LEAST SQUARE REGRESSION

Only when seven factors are extracted from the thirteen predictors it accounts for 99.97% of the total variation (Table 6). The estimates of the regression coefficient are given in Table 5.

TABLE 4 ANALYSIS OF VARIANCE FOR CLOSING PRICES

Source	Sum of Squares	Df	Mean Square	F	Sig.
Reg.	7.728E+09	7	1.104E+009	748078.6	0.00
Error	1058127.34	717	1475.770		

TABLE 5 ESTIMATE OF REGRESSION COEFFICIENT

	Estimate	Standard Error
Constant	-3703.264	472.053
OPEN	0.128	0.003
GH	0.140	0.002
LOW	0.146	0.002
VOLUME	0.000	0.000

	Estimate	Standard Error
ADJ_CLOSE	0.152	0.002
SO	174.192	7.208
ROC	-3386.024	475.739
MOMENTUM	0.214	0.031
MA5	0.146	0.001
MA7	0.144	0.001
MA10	0.143	0.001
DISPARTY	7021.329	328.459
PRICE_OSCILLATOR	2386.821	201.388

TABLE 6 PERCENT VARIATION EXPLAINED BY FACTORS FOR PREDICTORS AND RESPONSES

Factors	Variation Explained for Predictor(s)		Variation Explained for Response(s)	
	Percentage	Cum. Percentage	Percentage	Cum. Percentage
1	57.332	57.332	99.492	99.492
2	6.499	63.831	0.437	99.930
3	14.085	77.916	0.043	99.973
4	12.483	90.399	0.010	99.983
5	5.506	95.904	0.001	99.984
6	3.807	99.712	0.001	99.985
7	0.260	99.971	0.002	99.986

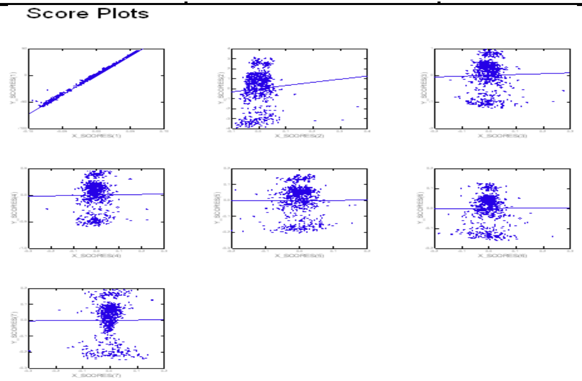


Figure 3 SCORE PLOT

The value of the estimates and the ANOVA table for the PLS model are given in Tables 5 and 4 respectively. The Score plot in given in Figure 3.

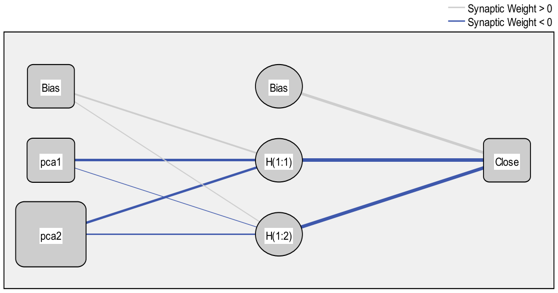
MULTILAYER PERCEPTRON

Taking the principal components as covariates and the closing prices as the dependent variable different networks are tried by rescaling the covariates and the dependent variables to standardize, normalize and adjusted normalize.

TABLE 7 PARAMETER ESTIMATES

		Predicted		
		Hidden Layer 1		Output Layer
Predictor		H(1:1)	H(1:2)	Close
Input Layer	(Bias)	.298	.047	
	pca1	-.402	-.032	
	pca2	-.481	-.221	
Hidden Layer 1	(Bias)			.542
	H(1:1)			-.551
	H(1:2)			-.850

The activation function of the output and the hidden layers are also changed (sigmoid, hyperbolic tangent and identity). The sum of square error and the relative errors are least only when the dependent variables are normalized and the covariates are adjusted normalized with the activation function of the hidden layer is hyperbolic tangent and that of the output layer is identity.



Hidden layer activation function: Hyperbolic tangent

Output layer activation function: Identity

Figure 4 NETWORK DIAGRAM

Network diagram is given in Figure 4. Root Mean Square Error and the Mean Average Percentage Error values are calculated for all the models proposed. The values are given in Table 8. The last fifty predicted values of all models with the actual closing prices in given in Figure 5.

TABLE 8 ERROR MEASURE

Model	RMSE	MAPE

MLR	141.615	0.007552
PLS	38.22171	0.001956
MLP	158.1483	0.008583

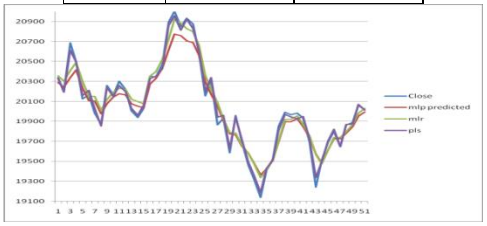


Figure 5 ACTUAL VS PREDICTED VALUES

CONCLUSION

Multilayer Perceptron model have predicted the daily closing prices with minimum MAPE value. But the PLS regression surpassed this model based on the Root Mean Square value and MAPE value. The RMSE value for PLS is 38.22171 while that of MLP 158.1483. Thus PLS regression with seven factors extracted from it is the best model in predicting the highly volatile closing prices.

REFERENCES

1. D. R. Cox and N. J. H. Small, "Testing Multivariate Normality", Biometrika, Volume 65, pp. 263-272.
2. D. R. Cox and Wermuth, "Tests of linearity, Multivariate Normality and the adequacy of linear scores", Applied Statistics, Volume 43, pp. 347-355.
3. J. Galindo, " A framework for comparative analysis of statistical and machine learning methods: An application to the black scholes option pricing equations", Techincal report Banco de Mexico, Mexico, DF,04930, 1998.
4. J. M. Hutchinson , Loaw and T. Poggio, "A nonparametic approach to pricing and hedging derivative securities via learning networks",Journal of Finance, 1994, Volume 49, pp. 851–889.
5. J. Yao and C. L. Tan, "A case study on neural networks to perform technical forecasting of forex", Neurocomputing , Volume 34, 2000, pp. 79-98.
6. K. V. Mardia, "Measures of multivariate Skewness and Kurtosis with applications", Biometrika, Volume 57, 1970, pp. 519-530.
7. K. V. Mardia, "Tests of univariate and Multivariate Normality", In Krishnaiah P.R. (ed.), Hand book of Statistics, Volume 1, Chapter 9, Amsterdam, North Holland, 1980.
8. Lean Yu, Shouyang Wang and Kin Keung Lai,"Mining Stock Market Tendency Using GA Based Support Vector Machines", X. Dengan Y. Ye (eds): wine 2005, LNCS, 3828, Springer Verlag Heidelberg 2005, pp 336-345.
9. M. Steiner and H. Wittemper, "Portfolio optimization with a neural Network implementation of the coherent market hypothesis", European Jouranl of Operations Research, Volume 100, 1997, pp.27-40.
10. Mark O.Afolabi and Olatoyosi Olider, "Predicting Stock Prices Using a Hybrid Kohonen Self Organization Map", Proceeding of the 40th Hawaii International Conference on System Science, 2007.
11. Myungsook Klassen,"Investigation of some Technical Indexes of Stock Forecasting in Neural Network", World Academy of Science, Engineering and Technology, 5, 2005, pp. 75-79.

12. Pei Chan Chang and Chin Yuan Fan, "A Hybrid System Integrating a wavelet and TSK Fuzzy rules for stock Price Forecasting", IEEE Transactions on systems, Man and cybernetics,Part C, Applications and reviews, Vol. 38, No. 6, 2008, pp. 802-815.
13. Ralph B. D'Agostino, "An omnibus test of normality for moderate and large samples", Biometrika, Volume 58, 1971, pp. 341-348.
14. Samreen Fatima, and Ghulam Hussian, "Statistical models of KSE 100 index using Hybrid Financial Systems", Neuro computing. Elsevier Science Publishers B. V. Volume 71, Issue 13-17, 2008, pp 2742-2746.
15. T. B. Trafalis, H. Ince and T. Mishina, " Support Vector Regression in option pricing", Proceedings of Conference on Computational Intelligence and Financial Engineering , Hong Kong, 2003
16. W. Leigh, M. Paz and R. Purvis, "An analysis of a hybrid neural network and pattern recognition techniques for predicting shork-term increases in the nyse", Internation journal of Management Sciences, Volume 30, 2002, pp.69-76.
17. Xiaoping Yang, "Prediction of Stock Prices Based on PCA and BP Neural Networks", Chinese Business Review, vol. 4, No. 5, S. No. 23, 2005, pp 64-68.