Physica A 438 (2015) 625–633

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Stock price forecasting for companies listed on Tehran stock exchange using multivariate adaptive regression splines model and semi-parametric splines technique

Mohammad Mahdi Rounaghi [a,*], Mohammad Reza Abbaszadeh [b], Mohammad Arashi [c]

[a] Department of Accounting, Mashhad Branch, Islamic Azad University, Mashhad, Iran

[b] Department of Accounting, Faculty of Economics and Business Administration, Ferdowsi University of Mashhad (FUM), Mashhad, Iran

[c] Department of Statistics, School of Mathematical Sciences, University of Shahrood, Shahrood, Iran

HIGHLIGHTS

- We used 40 variables for predicting stock price in Tehran Stock Exchange.
- MARS and semi-parametric splines predict stock price in Tehran Stock Exchange.
- Various comparison studies with different models exhibit superiority of our model.
- The proposed models provide very good results.

ARTICLE INFO

Article history:
Received 1 December 2014
Received in revised form 3 May 2015
Available online 28 July 2015

Keywords:
MARS model
Predicting
Stock price
Regression
Semi-parametric splines techniques

ABSTRACT

One of the most important topics of interest to investors is stock price changes. Investors whose goals are long term are sensitive to stock price and its changes and react to them. In this regard, we used multivariate adaptive regression splines (MARS) model and semi-parametric splines technique for predicting stock price in this study. The MARS model as a nonparametric method is an adaptive method for regression and it fits for problems with high dimensions and several variables. semi-parametric splines technique was used in this study. Smoothing splines is a nonparametric regression method. In this study, we used 40 variables (30 accounting variables and 10 economic variables) for predicting stock price using the MARS model and using semi-parametric splines technique. After investigating the models, we select 4 accounting variables (book value per share, predicted earnings per share, P/E ratio and risk) as influencing variables on predicting stock price using the MARS model. After fitting the semi-parametric splines technique, only 4 accounting variables (dividends, net EPS, EPS Forecast and P/E Ratio) were selected as variables effective in forecasting stock prices.

1. Introduction

Scholars are always looking for ways to predict the future. In this regard, techniques which have the lowest forecasting error are naturally viable and functional. So, for many years the mathematical methods such as simple average, weighted

* Corresponding author.
E-mail addresses: mahdi_rounaghi@yahoo.com (M.M. Rounaghi), abbas33@um.ac.ir (M.R. Abbaszadeh), m_arashi_stat@yahoo.com (M. Arashi).

http://dx.doi.org/10.1016/j.physa.2015.07.021
0378-4371/

average, double mean, regression, etc. were the only models used and confirmed firmly but these methods had some drawbacks in many cases as well. Developing artificial intelligence methods such as neural networks bring hopes and this continues that scientists consider them as a mathematical method replacement [1], particularly when we cannot make a relationship between the data and dependent and independent variables. It is anticipated that there is a complex relationship between the stock prices and macroeconomic variables on one hand and competitor or alternative asset prices in other markets on the other hand. Since there is no complete and comprehensive theoretical understanding about the mentioned variables and their relationships, using MARS method for modeling these relationships is a useful approach. Predicting stock prices at stock exchange is one of the most challenging issues in this field and it is found interesting by the investors and market activists. Not only these people care about the stock price forecasting but also it is interesting for many researchers and scholars. These markets are even more interesting when they are complex and unstable and because many variables are influencing these kinds of markets, the researchers look out new methods for predicting these markets.

The optimal allocation of resources is one of the most basic economic issues that have concerned natural and legal persons, economic decision-makers and officials involved in the capital market. Resource allocation is possible when resources are directed toward high returns investments with rational risk. The important role of the processed data that are combined with a valid measure is quite evident here. Since the capital market plays an important role in providing financial resources, directing capitals and promoting investments, measuring the effectiveness of stimuli with high predictive power for active decision-making process is one of the most important issues in capital markets in many countries. Hence, if capital markets are stable or improving, it can be concluded that the overall economy is growing. Therefore, more attention must be paid to the market and to the fundamentals of decision-making in it.

Market analysis that is sometimes known as technical analysis method is based on three basic and broad principles. The three principles that apply in all world markets are: (1) All information about a share is reflected in its price. Technicians believe that all knowledge available about the shares, whether economic, political, or psychological knowledge, is manifested in the stock market price. Unlike fundamental analysts, some furious technicians argue that efforts to study data and financial statements of companies and their revenue and supply and demand factors are in vain. They believe that all the information appears in the "stock price". (2) The second principle states that prices move in specific trends or processes which resist abrupt changes. Supply and demand for a product allows the share to move in a state of equilibrium. When a movement begins, no change will be sustainable in it unless it is over. For example, if the share price started to rise, it will continue to increase until they reach a certain point of return. (3) The third principle relates to the repetitive nature of market activity. According to this principle, the market is constantly repeating itself. Therefore, certain patterns can be found at various intervals on the chart. Of course, this is a psychological principle that people in similar conditions show similar responses. Since the capital market is a reflection of human performance, the technicians investigate the reactions to anticipate similar responses in similar situations [2].

One of the most important reasons for companies to look out for methods to predict the stock price changes, is the factors which influence the stock price changes that are unknown. Stock prices forecasting or stock return is possible through discovering the behavior patterns of the stock prices generating process and success rate of the discovery of these behavior patterns which show forecasting efficiency. In fact, we could investigate the stock price generating process as a dynamic pattern. The mentioned process could be obtained by linear models, nonlinear models or stochastic models.

2. Literature review and theoretical basis

In the early twentieth century, some security market participants believed that price historical reviewing provide useful information for predicting prices at future so they believe if they could obtain price trends, they could discover change patterns and it would tell us when certain trends will occur. These people are called Chartists because they focus on charts. They believed that they could identify and predict supply and demand relationships as a result of stable trends for any stock or whole market. More importantly, technical analysts believe that investors behave in a predicted way when they face similar situations that occurred in the past. In other words, the history repeats Rajabi divarzam et al. [3].

Stock price forecasting

Stock price is an index which represents the general level of stock prices of companies listed in Stock Exchange. Investigating time series patterns of a company stock price is its prices at opening and closing market or its highest or lowest values for each day and many days. Tradable financial assets are the basis of market decisions. Investors are looking at the latest news of company's stock prices and analyze for future fluctuations forecasting. Since analyzing markets through prices is easy for everyone, so it is more common to use stock price.

For predicting stock prices and finding the optimum stocks, there are two groups of experts. The first group is called the fundamental analysts; they try to predict stock prices by domestic stock price and investigate the factors influencing the stock prices. In this method, for accurate stock price determination, we identify and investigate all factors influencing stock price. Determining stock value and comparing it with its price in exchange market, we select stocks which are exchanging at a price lower than their real value to buy and stocks which are exchanging at a higher price to sell.

The second group is called technical analysts; they believe that they could predict stock performance in future by investigating its behavior at past. In this method, we investigate stock price to find good opportunities for selling and buying.

Analysts assume that all factors influencing the stock value are reflected in its rate. In other words, when prices are set based on supply and demand, the influencing factors such as economical and political conditions governing market, market participant evaluation about stock future, manager role and his strategies for company, inflation, etc. are all influenced the price. In addition, analysts assume that stock price changes follow predictable patterns and trends. In other word, reviewing stock price at the past, we could obtain valuable information to predict price at future [4]. There are several studies for predicting stock price and we review some of them below.

Olson and Mossman [5] also found that artificial neural network method is more accurate than other methods for predicting stock price of companies listed in Canada stock. Yang's (2009) studies about stock price index at Taiwan stock exchange showed that the data fluctuations are consistent with ARCH patterns and combining it with artificial neural network could contribute to forecasting improvement. Roh's [6] findings on Korea Stock price index showed the priority of hybrid artificial neural network model and time series model for predicting.

Ravazzola and Phylaktis [7] showed that there is a long term bivariate causality between stock prices and real exchange rates for Pacific Ocean countries after inserting US stock index variable (as a channel between currency market and stock market). Mohammad Nashat and Mostafa Khalid [8] showed that stock price influenced the exchange rates significantly and negatively in long term and there is a one way causality. Also they concluded that Pakistan stock exchange is inefficient. Marcek [9] presented Jenkins-Banks analysis in time series. Then they explained autoregressive modeling for stock price forecasting and described the fuzzy autoregressive model and fuzzy-neural network as two alternative methods for autoregressive model of stock price forecasting. Wang [10] used "grey prediction fuzzy system" technique in its studies to predict stock price at Taiwan Exchange. The main problem for stock forecasting is data integrity and its high volume and to resolve this problem they used "data center" technique. Koulouriotis et al. [11] used forecasting techniques such as multivariate regression, fuzzy-neural networks and neural inference-fuzzy adaptive systems and conducted a comprehensive study about this techniques' application for short-term stock price forecasting. Zapranis and Fransis (1994) modeled stock price behavior using neural networks and compared their performance with regression models. In this study, they used neural networks as an alternative to classic statics techniques to forecast large companies' stock prices. The results proved that neural networks work better than statistic techniques and present better models. Saad et al. [12] designed a system that forecast short term stock price changes. First there were preprocessing of data and then neural network is modeled to estimate opportunities for profit making.

Kuo, Chen and Hwang [13] reported a consulting system to keep, sell or buy stocks, in an article "An Intelligent Stock Trading Decision Support System Through Integration of Genetic Algorithm Based Fuzzy Neural Network and Artificial Neural Network, Fuzzy sets and systems". One of the main features of this system is that it allows its users to quantify quality variables influencing stock price forecasting. These researchers presented a similar article in 1998 regardless of genetic algorithms. In this article, a questionnaire with fuzzy-Delphi method was used for evaluating experts' opinion about stock price forecasting. Hashem Zare [14] has investigated the relationship between gold coins, property and currency variables as competing assets and cash volume, industrial production index, petroleum rate, and the inside and outside price ratio as some of the main macroeconomics variables influencing stock price index using ARDL model. The results showed that there is a positive relationship between stock price index and property price, petroleum price, gold coin price and the inside and outside price ratio. Also, there is a negative relationship between stock price index and exchange rate and cash volume. Mohamad Osolian [15] investigated the relationship between exchange rate, price of crude oil per barrel and inflation rate and also its effects on stock price for companies listed in stock exchange. The result has shown that there is no significant relationship between these variables and companies' stock price. Mohsen Mokhtari [16] proved variance analysis and analysis of co integration of positive effect of money on stock price and capital cost reduction theory, using VAR model. Mostafa Karimzade [17] investigated the relationship between monetary variables and stock price index using ARDL model. In this study there is a positive relationship between stock price index and liquidity and a negative significant relationship between stock price index and real exchange rate and interest rates. Toloui and Haghdost (2007) had written an article "stock price forecasting modeling using neural networks and comparing it with mathematical predictions". In their article, they investigated neural network and regression models' performance for stock price forecasting data and measured prediction error of these two methods. The results confirmed the precision and low error of intelligent neural model.

Semi-parametric models were first introduced by Engle and his colleagues in 1986. Engle et al. [18] and Chenv and Xiao (1991) using the least squares penalties method (Spline smoothers), Fridman [19] using spline multivariate adaptive regression (MARS), Cuzick et al. [20] using minor residuals, Surin and Wang [21] using profile likelihood evaluated this model parameters. Ashmalnsy and Stalker (1999) used these models for investigating the relationship between household diesel consumption and income, age, number of family members, residence, and house type variable in US. They found that there is non-linear relationship between the diesel consumption logarithm per barrel and the age and income variable and there is a linear relationship between the number of family members and people who are driving and other variables.

3. Research method and research variables

The population of this research includes all firms listed in Tehran Stock Exchange from 2006 until the end of 2012. All the cases in this study were selected by using screening method (such as the Cochrane statistical sampling techniques)

Table 1
Research variables.

Independent variable (Accounting variables)	Book Value, Net EPS, forecasted EPS, ROE, ROA, P/E, Percentage of dividends, dividends to price, the Beta, CFO, operating profit (Risk), profit in cash, dividend growth, rating liquidity, firm size, book value to market value, market value to book value, the sum of shareholders' equity at the end fiscal year, net assets, annual return of shares, capital raise, changes in the price index of Tehran Stock Exchange, Tehran Stock Exchange trading volume (million shares), industry index, financial index, weighted index of 50 active companies, index of dividend yields, the value of stock (trillion Rials), the market value (trillion Rials)
Independent variable (Economic variables)	GDP, the average price of crude oil per barrel (USD), bank interest rate (percent), bank deposit rates (percent), inflation Gold rate (Rial), Average exchange rate (Toman), Economic growth (percent) Liquidity (thousands trillion Tomans), National Savings (trillion Rials)
Dependent variables	The final stock price

considering the below criteria:

1. Full information about each studied company within research time scope should be presented.
2. Companies should not change their fiscal year within research time scope.

3.1. MARS

In this study, we used MARS model and semi-parametric splines technique. MARS model as a nonparametric method is an adaptive regression method and it is working well for problems with high dimensions and high predictive variables. Modeling in this method is based on a segment linear regression fitting which is the simplest type of splines. Based on MARS method, the influential variables and endpoints of intervals (nodes) are identified for each variable through a quick but very intense way. MARS model in addition to searching variables one by one, seeks interaction effects between variables until the optimum stage. MARS optimal model is selected in a two-stage process. Firstly, MARS creates a model which is too big through a formal mechanism and secondly it removes basic functions that have the smallest contribution to reach an optimal model.

3.2. Semi-parametric splines technique

Smoothing splines is a nonparametric method. One of the strong points of spline fitting is their comfortable estimation in semi-parametric models. This model will have a better performance if there are many predictive variables. Splines are piecewise regression functions that are connected to each other by points that are called knots. In principle, separate regression equations are fitted within the area between any two knots and these pieces connect each other in these parts. In fitting the model by splines, the analyst must determine the degree of polynomial and the number and the location of knots before. In this study for predicting the stock price we used 40 variables (30 accounting variables and 10 economic variables) as described in Table 1.

4. MARS

MARS method can be considered as a generalization of stepwise linear regression or modification of the tree regression (CART) [22]. In this article, we introduce MARS from the generalization of the stepwise linear regression point of view.

If the optimality criterion for the regression model is the coefficient of determination (R2), we select in the MARS method by examining potentially large number of nodes, the node that has the highest coefficient of determination [23]. Although finding the best pairs of nodes need much computation and also when the number of required nodes are unknown, finding the best set of nodes is difficult and exhausting. MARS finds the location and number of required nodes in a forward–backward way. In the forward stage, the model which includes over estimation generates too many nodes. Then, those nodes which have the minimum contribution to the overall fitting have to be removed. So, selecting nodes at forward stage includes several incorrect places for nodes but these incorrect nodes will be removed at backward stage. So we could call MARS method as adaptive spline method.

The basis of MARS is dependent on segment functions which called spline functions and they are defined as below:

$$h_1(x) = (t - x)_+ = \begin{cases} t - x & \text{if } t > x \\ 0, & \text{Otherwise,} \end{cases}$$

$$h_2(x) = (x - t)_+ = \begin{cases} x - t & \text{if } x > t \\ 0, & \text{Otherwise.} \end{cases}$$

Developing strategy for this model is like a stepwise linear regression. However, we use functions in collection C and its multiple (interaction effects) instead of using input variables.

$$C = \left\{(X_i - t)_+,\ (t - X_j)_+\right\}, \quad t \in \left\{x_{1j}, x_{2j}, \ldots, x_{nj}\right\},\ j = 1, \ldots, p.$$

If there is no repeated data in all the data, collection C includes 2np basic functions. The MARS model has the following structure: (1-4)

$$f(x) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(X)$$

in which, each $h_m(x)$ is a function from C or multiple two or several functions from C. Also M is the number of functions in the model which will be identified after forward stage. Having h_m functions, we could estimate β_m through minimum sum of squares error.

First, we describe MARS only with one input variable x, and then we generalize it to multiple inputs. In this situation, we consider observed data as $((x_1, y_1), \ldots, (x_n, \ldots, y_n))$ where (x_i, y_i) is representing ith observation.

At early stage, this model only includes y-intercept, that is:

$$\hat{f}_1(x) = \hat{\beta}_0.$$

When $\hat{\beta}_0 = \bar{y}$ this model is called model 1, then for creating model 2, we select the model which includes the minimum sum of squares error among other models.

$$\hat{f}_{21}(X) = \hat{\beta}_0 + \hat{\beta}_1(X - x_1)_+ + \hat{\beta}_2(x_1 - X)_+,$$
$$\hat{f}_{22}(X) = \hat{\beta}_0 + \hat{\beta}_3(X - x_2)_+ + \hat{\beta}_4(x_2 - X)_+,$$
$$\cdots\cdots$$
$$\hat{f}_{2n}(X) = \hat{\beta}_0 + \hat{\beta}_{2n-1}(X - x_n)_+ + \hat{\beta}_{2n}(x_2 - X)_+.$$

Assume that we select a model as below:

$$\hat{f}_2(X) = \hat{\beta}_0 + \hat{\beta}_3(X - x_2)_+ + \hat{\beta}_4(x_2 - X)_+,$$

which can be presented briefly as below and call it model 2:

$$\hat{f}_2(X) = \hat{\beta}_0 + \hat{\beta}_3 h_3(X) + \hat{\beta}_4 h_4(X).$$

Now, we develop model 3 with adding other basic functions to model 2, adding those functions reduce sum of squares error too much. Therefore, model 3 will be chosen from the following models:

$$\hat{f}_{31}(X) = \hat{\beta}_0 + \hat{\beta}_1^* h_3(X) + \hat{\beta}_2^* h_4(X) + \hat{\beta}_3^* h_1(X) + + \hat{\beta}_4^* h_2(X) +,$$
$$\hat{f}_{32}(X) = \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots + \hat{\beta}_3^* h_5(X) + + \hat{\beta}_4^* h_6(X) +,$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$
$$\hat{f}_{3(n-1)}(X) = \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots + \hat{\beta}_3^* h_{2n-1}(X) + + \hat{\beta}_4^* h_{2n}(X).$$

As it is said, a model that reduces $\sum_{i=1}^{n} \left(y_i - \hat{f}(x_i)\right) 2$ much more than others will be selected. It should be said that every time when we develop a new model, the new model coefficient β will be estimated through minimum sum of squares again. This process is continued until we add k basic functions to model. Note that $k \leq 2n$ or it is identified by user.

At the end of forward stage, the developed model is a model that although it is possible for modeling data but it causes data overestimation. For solving this problem, MARS method will go to the next stage that is "removing basic functions from the model". Therefore, we consider the possible removal of each f_k function whose y-intercept is one and $2(k-1)$ is a basic function.

Those basic functions will be removed that their removal increase the sum of squares error as less as possible. This process continues until all functions except y-intercept is removed from the model.

When the removal process of basic function is done completely, we develop $2k-2$ model (in each stage one model) that each of them is a candidate for the final model. We calculate generalized cross validation (GCV) criterion for each of these models. If GCV_L is the GCV for the Lth model in the backward removal for $L = 1, \ldots, 2K-2$ so we have:

$$GCV_L = \frac{SSE_L}{1 - ((VML + 1)/n)}$$

where m_1 and SSE_L are number of basic functions and sum of squares error at Lth respectively and V is the penalty for each basic function. In fact, we could say that V is a smoother variable that user defines and it controls the transaction between simple and complex models and it is between 2 and 4. The model that has the lowest GCV among $2K-2$ will be selected as the final model and an estimation for MARS.

If there is more than one prediction variable, MARS could insert their multiplying that is international effect in model except for basic function.

$$h_m(X) = h_i(X)(X_j - t)_+, \qquad h_{m+1}(X) = h_i(X)(t - X_j)_+$$

where $\left\{(X_j - t)_+, (t - X_j)_+\right\}$ is reflected pairs from C collection for $j = 1, \ldots, p$ and $(1 \le i \le m-1)$ $h_i(X)$ is basic function that already existed in model. Selecting pair functions of $\{h_m(x), h_{m+1}(X)\}$ is done through searching multiplication of basic function $h_i(x)$ from model and reflected pairs at C collection so that $h_i(x)$ and reflected pairs do not have identical variable X_j in their structure and when they add to the model, have the most reduction for some of squares error [24],

First for clarifying issue, assuming two variable X_1 and X_2 and observing; $i = 1, \ldots, n$ $x_i = (x_{i1}, x_{i2})$, the processed MARS model can be presented as below:

$$f(X) = \beta_0 + \left(\sum_{i=1}^{n} \beta_{i1} h_{i1}(x_{i1}) + \sum_{i=1}^{n} \beta_{i2} h_{i2}(x_{i2})\right) + \sum_{j=1}^{n}\sum_{i=1}^{n} \beta_{(ij)12} h_{i1}(x_{i1}) h_{i2}(x_{j2})$$

Dual effect + (main effect) + y-intercept (2-4)

where h_{i1} and h_{i2} are basic functions based on variable X_1 and X_2.

As generalization of model (2-4) when we have p as prediction variable, the MARS model is as below:

$$f(x) = \beta_0 + \sum_{j=1}^{p}\left(\sum_{i=1}^{n} \beta_{ij} h_{ij}(x_j)\right) + \sum_{j=1}^{p}\sum_{k>j}^{p}\left(\sum_{i=1}^{n}\sum_{l=1}^{n} \beta_{iljk} h_{ij}(x_j)\, h_{lk}(x_k)\right) + \cdots + \sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\cdots\sum_{i_p=1}^{n} \beta_{i_{1,\ldots,i_p}} h_{1,i_1}(x_1)\, h_{2,i_2}(x_2) \cdots h_{2,i_p}(x_p) \quad .$$

Higher effects + dual effects (binary interactional) + original effects + y-intercept.

It is not necessary that model includes all interactional effects and so determining its rank is up to user. Often an additive model as below (original effects + y-intercept) is enough:

$$f(x) = \beta_0 + \sum_{j=1}^{p}\left(\sum_{i=1}^{n} \beta_{ij} h_{ij}(x_j)\right).$$

In general, we could consider spline model which includes dual interactional effects as an approximation for f as follows:

$$f(x) = \beta_0 + \sum_{j=1}^{p}\left(\sum_{i=1}^{n} \beta_{ij} h_{ij}(x_j)\right) + \sum_{j=1}^{p}\sum_{k>j}^{p}\left(\sum_{i=1}^{n}\sum_{l=1}^{n} \beta_{iljk} h_{ij}(x_j)\, h_{lk}(x_k)\right).$$

Models with triplex interactional effects or even with higher rank interactional effects also could be considered but they are not very common [25].

5. Semi-parametric splines

Splines are piecewise regression functions that are connected together in points that are called nodes. In principle, separate regression functions are fitted within the area between any two nodes and pieces are connected to each other in these nodes. One of the advantages of splines-based fitting is its ease of estimation in semi-parametric models (Luke Kiel, 2008). To fit the model by using splines, the analyst must determine the degree of the polynomial and the number and locations of nodes before.

Perhaps the most confusing aspect of splines is that there are so many different types of them. Some of these are regression splines, cubic splines, p-splines, natural splines, splines and thin plate splines and smoothing splines. Moreover, their synthesis is also used as a spline, for example natural cubic splines.

Selecting bounds, areas and distances in splines are very important. A node specifies the beginning of one area and the end of the other. Usually nodes are created where the behavior changes. Two fundamental points are noticed while selecting nodes:

1. The accurate selection of areas with regard to their bounds.
2. The correct determination of the number of intervals required for each variable (i.e., if a function is very irregular in an area, large distances are required. In contrast, if the function is linear in an area, only one space is needed).

Estimation in complex semi-parametric regression model as an additive model using smoothing splines

Since the semi-parametric regression model is a special case of additive models, additive estimation methods are used for estimating purposes. The most common method for the estimation of additive models is to estimate each function by an arbitrary smoother. Some of these smoothers include smoothing splines, locally weighted smoothers, kernel smoother and running lines.

In this paper, the smoothing spline is used in the same manner as other simple semi-parametric models. Hence the criteria of least squares must be defined for these models. Before defining the criteria, it must be determined how to define the incidence matrix in these models.

Suppose that some of the observed values $X_j;\ j = q+1, \ldots, p$ are repeated in the following semi-parametric model:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_q X_q + f_{q+1}\left(X_{q+1}\right) + \cdots + f_p\left(X_p\right) + \varepsilon.$$

Similar to simple semi-parametric model

$$y_i = x_i'\beta + f\left(t_i\right) + \varepsilon_i, \quad i = 1, \ldots, n$$

incidence matrix is needed to obtain a suitable method for sorting the observable values of variables X_j. To this end, suppose that distinct and regular values $X_{j1}, \ldots, X_{jn}$ for the jth predictive variable are specified by $s_{j1}, \ldots, s_{jq},\ j = q+1, \ldots, p$. Thus, the observed distinct values for each variable X_j can be displayed as a vector s_j. The relationship between $X_{j1}, \ldots, X_{jn}$ and $s_{j1}, \ldots, s_{jq}$ is demonstrated by matrix N_j of dimension $n \times q$, so that

$$N_j = \begin{cases} 1 & x_{ji} = s_{jk} \\ 0 & \text{otherwise} \end{cases}, \quad i = 1, \ldots, n,\ k = 1, \ldots, q,\ j = q+1, \ldots, p.$$

Therefore, the simple semi-parametric model can be rewritten as follows:

$$Y = X\beta + \sum_{j=q+1}^{p} N_j f_i\left(S_j\right) + \varepsilon \tag{1-5}$$

where $X = (x_1, x_2, \ldots, x_n)',\ i = 1, \ldots, n$ is active q-dimensional vector, $\beta = (\beta_1, \beta_2, \ldots, \beta_q)'$ is passive q-dimensional vector, $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)'$ is random error vector, and N_j is the incidence matrix corresponding to every variable X_j for $j = q+1, \ldots, p$. Also f_i are smooth passive functions that should be estimated and S_j are the vectors of observed regular and distinct values of X_j for $j = q+1, \ldots, p$.

To develop what has been said about the smoothing spline criteria for simple semi-parametric models, the penalized least squares are generalized for complex semi-parametric criteria as follows:

$$S\left(\beta, f\right) = \left\{Y - X\beta - \sum_{j=q+1}^{p} N_j f_i\left(S_j\right)\right\}^2 + \sum_{j=q+1}^{p} \alpha_j \int f_j''^2\left(t\right) \mathrm{d}t.$$

Notice that every function in the above expression is penalized by a distinct constant α_j. In the following theorem due to Valizadeh et al. [26], estimates f and β in the semi-parametric model (1-4) are obtained by minimizing the penalized least squares criterion $S\left(\beta, f\right)$.

Theorem. *Under the assumptions of this section, the least squares estimates β and f_k for $k = q+1, \ldots, p$ for semi-parametric model (1-5) can be obtained from the following equations:*

$$N_k\hat{f}_k = S_k\left(Y - X\hat{\beta} - \sum_{\substack{j=q+1 \\ j\neq k}}^{p} N_j\hat{f}_i\right)$$

$$\hat{\beta} = \left(X'X\right)^{-1}X'\left(Y - \sum_{j=q+1}^{p} N_j\hat{f}_i\right)$$

where $M_k = (I - S_k)$, $S_k = N_k(N_k'N_k + \alpha_k K_k)^{-1}N_k'$ and α_k are smoothing parameters and K_k is a positive symmetric matrix.

6. Results

6.1. Result of MARS model

For fitting MARS model into two key parameters, some basic functions (nk) should be optimized at forward stage and interactional effects degree (deg). In this way, changing basic functions nk = 11, 12, ..., 50 deg = 1, 2, various models are fitted for data that among them a model with setting nk = 13 deg = 2 is an optimum model, which includes maximum determination coefficient and minimum GCV and RSS. In our MARS model for this data, out of 40 variables, 4 of them are incorporated.

Final MARS model with above setting has the form:

$$\begin{aligned} y = {} & 847.8019 + 16.4486h\left(X_3 - 36\right) + 0.2688h\left(36 - X_3\right) + 35.9955h\left(X_6 - 26.02\right) \\ & - 23.2926h\left(26.02 - X_6\right) + 9803.2711h\left(-1.31 - X_9\right) - 2.4617h\left(X_1 - 5359.01\right) * h\left(-1.31 - X_9\right) \\ & - 2.0912h\left(5359.01 - X_1\right)h\left(-1.31 - X_9\right) + 1.0183h\left(X_3 - 36\right) \\ & * h\left(X_6 - 16.82\right) - 0.9772h\left(X_3 - 36\right) * h\left(X_3 - 36\right) - 0.9942h\left(36 - X_3\right) \\ & * h\left(X_6 - 0.28\right) + 0.887h\left(36 - X_3\right) * h\left(-0.28 - X_6\right) \end{aligned}$$

Table 2
The coefficient of determination values.

The total squared error (RSS)	GCV	Determination values
2 430 346 246	952 309.2	0.9735

Table 3
Relationship between variables.

Variables	Spearman correlation coefficients
Net EP and final stock price	0.575
EPS forecast and final stock price	0.775
P/E and final stock price	0.502
Dividends and final stock price	0.543

Table 4
Coefficient of determination.

The total squared error (RSS)	The coefficient of determination
23 749 199 256	0.743

where X_1 is book value per share, X_3 is earnings prediction for each share (FEPS), X_6 is price to earnings (P/E), X_9 is risk and Y is final price per share. Determination coefficient values GCV and RSS for above model are presented in Table 2.

The value 0.9735 shows the appropriateness of fitting the model for the data.

6.2. Result of semi-parametric splines

According to the result of the Kolmogorov–Smirnov test that checked the normality of the variables and based on the significance level of variables (0.000), it was found that none of the studied variables follow a normal distribution.

Most tests and tools used in statistics assume that any error in a financial model are Gaussian distributed, and it is a common practice in economics to use a Gaussian distribution to approximate empirical data. Mandelbrot and Fama were among the first to notice that the logarithm of cotton price fluctuations and common stock price fluctuations have fatter tails than those produced by a Gaussian distribution, and they proposed a stable Lévy distribution to model the stochastic properties of the fluctuations. Analyzing high-frequency data, Mantegna and Stanley reported that the stable Lévy distribution accurately models only a broad central region of the probability distribution function (PDF) of stock price changes, whereas Gopikrishnan et al. reported that a power law with an exponent value beyond the Lévy regime is needed to describe the tails [27].

Hence, Spearman nonparametric correlation coefficient was used to evaluate the effect (correlation coefficient) of explanatory variables on the response variable. Then explanatory variables, that have 50% relationship with the response variable, were exploited to be able to present in the regression model. Among the studied variables, net EPS, EPS forecast, P/E ratio and dividends were associated with the response variable by more than 50%. The values of these relationships are shown in Table 3.

The following semi-parametric regression model can be suggested for the study of variables on the basis of smoothing splines.

$$y = \beta_0 + \beta_1 X_1 + f_1(X_2) + f_2(X_3) + f_3(X_4)$$

where X_1 is dividend, X_2 is net EPS, X_3 is EPS forecast, X_4 is p/E ratio, y is final price, $\beta_1 X_1$ is the parametric component of the model and $f_1(X_2)$, $f_2(X_3)$, $f_3(X_4)$ are nonparametric components of the model among which f_1, f_2 and f_3 are estimated based on smoothing splines.

The sum of squared error of this model and its coefficient of determination are presented in Table 4.

7. Conclusion and recommendations

MARS model and semi-parametric splines technique can be used in econophysics research. Mariko and Isao, 2010 provide a nonparametric model of multi-step ahead forecasting in diffusion processes. The paper provides simulation studies to evaluate its performance of multi-step ahead forecasting by comparing with the global linear model, showing the better forecasting performance of the nonparametric model than the global linear model. The paper also conducts empirical analysis for forecasting using intraday data of the Japanese stock price index and the time series of heart rates. The result shows the performance of forecasting does not differ much in the Japanese stock price index, but that the nonparametric model shows significantly better performance in the analysis of the heart rates. Guanghui and Jianping, 2008 used a nonparametric approach for European option valuation. A nonparametric approach for European option valuation is

proposed in this paper, which adopts a purely jump model to describe the price dynamics of the underlying asset, and the minimal entropy martingale measure for those jumps is used as the pricing measure of this market. A simple Monte Carlo simulation method is proposed to calculate the price of derivatives under this risk neural measure. And the volatility of the spot market can be renewed automatically without particular specification in the proposed method. The performances of the proposed method are compared to that of the Black–Scholes formula in an artificial world and the real world. The results of our investigations suggest that the proposed method is a valuable method.

MARS model can work using real specified values for Tehran Stock prices with more than two variables. This research recommends that the specified MARS model can be used for clarifying efficient process for presenting proper methods for modeling and its complex series volatile. On the other hand, we could investigate as a technical recommendation, the capability of genetics algorithm model and non-linear regression models and fractal models for predicting stock price as common non-linear regression models. Furthermore, the application of semi-parametric splines technique for time series is being verified and it may also improve the predictions. One of the implications of this research is that it explains an efficient process in order to provide an appropriate way of modeling and predicting complex and volatile series.

Acknowledgments

The authors would like to thank Ms Valizadeh (Amirkabir University, IRAN) for her help.

The authors would like to thank Dear professor H. Eugene Stanley, the main editor, and two anonymous reviewers for many valuable comments and suggestions.

References

[1] Abbas Toloiashlqi, Shadi Haqdost, Stock price modeling using neural networks and comparing them with mathematical prediction methods, Econ. J. (24) (2006) 237–252. Science-research.
[2] P. Piri, The effect of market structure indicators on the stock prices of companies listed in Tehran Stock Exchange, Q. J. Secur. Exch. 19 (2012) 27–41.
[3] Rajabi Divarzam Farzin, Alireza Darzi, Non linear modeling and behavior prediction of stock price for cement companies listed in Tehran Stock Exchange, Chapter 1, No. 63, 17–26, 2013.
[4] Hamed Meihaninia, Prediction of stock price using unobserved components model with stochastic volatility, financial management. Master of Science Essay, Engineering Faculty, Science and Culture University, 2012.
[5] D. Olson, C. Mossman, Neural network of Canadian stock returns using accounting ratios, Int. J. Forecast. 19 (2003) 453–465.
[6] T.H. Roh, Forecasting the volatility of stock price index, Expert Syst. Appl. 33 (2007) 916–922.
[7] B. Ravazzola, C. Phylaktis, A bivariate causality between stock prices and exchange rates: Evidence from recent Asian flu, Quart. Rev. Econ. Finance 40 (1998) 337–354.
[8] N. Mohammad, K. Mostafa, Exchang rate & stock prices relationship: Empirical evidence from Pakistan financial markets, J. Finance 45 (2002) 1237–1254.
[9] Dusan Marcek, Stock price forecasting: Autoregressive modelling and fuzzy neural network, Mathw. Soft Comput. (7) (2002) 139–148.
[10] Yi-Fan Wang, Predicting stock price using fuzzy grey prediction system, Expert Syst. Appl. 22 (2002) 33–38.
[11] Dimitris Koulouriotis, Dimitris Emiris, Ioannis Diakoulakis, Constantin Zopounidis, Behavioristic analysis and comparative evaluation of intelligent methodologies for short-term stock price forecasting, Fuzzy Econ. Rev. (2) (2002) 23–57.
[12] E. Saad, D. Prokhorov, D. Wunsch, Advanced neural-network training methods for low false alarm stock trend prediction, in: Proc. IEEE Int. Conf. Neural Networks, Washington, DC, June, 1996.
[13] R.J. Kuo, C.H. Chen, Y.C. Hwang, An intelligent stock trading decision support system through integration of genetical gorithm based fuzzy neural network and artificial neural network, Fuzzy Sets and Systems 118 (1) (2001) 21–45.
[14] Hashem Zare, A survey on competitor assets price effects and other macroeconomic variables on stock price index. Master of Science Essay, Shiraz University, 2005.
[15] Mohammad Osolian, A survey on macroeconomics variables effects on stock price index for companies listed on Tehran Stock Exchange between 1993–2002, Master of Science Essay, Management Faculty, Tehran University, 2005.
[16] Mohsen Mokhtari, Effects of monetary policy on stock returns in Tehran Stock Exchange, Master of Science Essay, Social and Human Science Faculty, Tabriz University, 2005.
[17] Mostafa Karimzade, A survey on long term stock price index relationship with macro monetary variables using collective techniques in Iran economy, Master of Science Essay, Beheshti University, 2004.
[18] R.F. Engle, C.W.J. Granger, J. Rice, A. Weiss, Semiparametric estimates of the relation between weather and electricity sales, J. Amer. Statist. Assoc. 81 (1986) 310–320.
[19] J.H. Friedman, Multivariate adaptive regression splines, Ann. Statist. 19 (1990) 11–41.
[20] J. Cuzick, Semiparametric additive regression, J. R. Stat. Soc. Ser. B 54 (1992) 831–843.
[21] T.A. Severini, W.H. Wang, Generalized profile likelihood andconditional parametric models, Ann. Statist. 20 (1992) 1768–1802.
[22] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning Data Mining, Inference, and Prediction, second ed., Springer, New York, 2009.
[23] Salford systeme:. MARS User Guide, 2001.
[24] M. Durmaz, M.O. Karsliglu, M. Nohutcu, Regginal VTEC modelingwith multivariate adaptive regression splines, Adv. Space Res. 46 (2010) 180–189.
[25] T. Valizade, Adaptive and non-adaptive splines in semiparametric regression models (M.Sc. dissertation), Shahrood University, Iran, 2012.
[26] T. Valizadeh, D. Shahsavani, M. Arashi, Improving efficiency by semiparametric multivariate adaptive regression splines. Submitted for publication, 2012.
[27] B. Podobnik, A. Valentincic, D. Horvatic, H.E. Stanley, Asymmetric Levy flight in financial ratios, Proc. Natl. Acad. Sci. USA 108 (44) (2011) 17883–17888.