Expert Systems with Applications 42 (2015) 6267–6275

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Stock market trend prediction using dynamical Bayesian factor graph

Lili Wang [a,*], Zitian Wang [b], Shuai Zhao [c], Shaohua Tan [c]

[a] *School of Economics and Management, Tsinghua University, Beijing 100084, China*
[b] *Agricultural Bank of China, Beijing 100005, China*
[c] *Center for Information Science, Peking University, Beijing 100871, China*

ARTICLE INFO

Article history:
Available online 20 March 2015

Keywords:
Dynamical Bayesian factor graph
Macroeconomic factors
Market trend prediction

ABSTRACT

A new qualitative method using the concept of dynamical Bayesian factor graph is developed in this paper for the prediction of stock market trend. The essence of this method is to compute the corresponding dynamical Bayesian factor graph for a selected set of macroeconomic factors over a period of time of interest. The computed time series of graphs capture both the mutual influential relationships and the evaluation of these relationships among these factors over a specified period of time. Then any topological structural change in the adjacent graphs at anytime predicts a change in market trend in a short future. Our computational analysis also indicates that if the topological structure of the underlying dynamical Bayesian factor graph is unchanged, the general market trend appears invariant. We demonstrate the effectiveness of this method by applying it in the analysis of market trends in the US stock market and Chinese stock market with good prediction results.

1. Introduction

Being able to predict a stock market index quantitatively, systematically and reliably is most desired by both financial theorists and practitioners. Decades of research have indicated, however, that this prediction problem remains to be elusive and difficult to tackle even with two currently well-known and widely accepted methodologies: The traditional risk-return modeling (see Fama & French, 1996; Naranjo, Nimalendran, & Ryngaert, 1998; Lettau & Ludvigson, 2001, 2005; Santos & Veronesi, 2006) and the pure mathematics inspired approach of financial time series forecast (Huang & Tsai, 2009; Huang, Nakamori, & Wang, 2005; Tay & Cao, 2001; Kim, 2003).

The reasons for often poor and, in many cases totally unacceptable, prediction performances in a practical sense using either of the methodologies have increasingly been better understood. Risk-return models, for instance, rely on market efficiency and selection of a sufficient set of factors to calculate intrinsic value of a target index as its prediction. It turns out, however, that intrinsic value is more an abstract theoretical number than a converging point of an actual market, thus is often totally unrelated to real market index value. Risk-return models are also negatively impacted by a possibly biased selection of factors for a model structure. Although a selection may be justified by various financial or economical views of some experts, the inherent structural bias by way of emphasizing on a few factors while ignoring other factors often leads to totally misleading prediction outcomes since the factors that are ignored in a prediction model may actually be dominating a predicted target at or around times of prediction.

The time series forecast methodology, on the other hand, is more of a mathematical modeling approach. The essence of this approach is to use a pre-selected mathematical model structure or a structural building algorithm and existing sample data to model an underlying stock market system. The computed model is then used to predict future sample data. In order for such an approach to work, a pre-selected model structure or a structural building algorithm will have to be able to adequately represent or capture the underlying stock market system. The earlier and relatively simplistic idea was to use a linear model structure for this purpose. The idea was quickly elevated to wide adoption of nonlinear model structure and learning algorithms due to the apparent nonlinear nature of the underlying stock market. Some of the techniques such as ANN (Artificial Neural Networks) and SVM (Support Vector Machines) with nonlinear kernel functions have gained popularity in a range of market prediction problems in recent years. Among the latest studies, Ticknor (2013) and de Oliveira, Nobre, and Zrate (2013) applied ANN to forecast closing prices of stocks. Żbikowski (2014) proposed volume weighted SVM to predict stock return and risk. Other more elaborate nonlinear modeling approaches have also been used. Patel, Shah, Thakkar, and

[*] Corresponding author. Tel.: +86 13601065576.
E-mail addresses: lili_wang2010@163.com (L. Wang), wzt@cis.pku.edu.cn (Z. Wang), zhaoshuai2013@pku.edu.cn (S. Zhao), tansh98@hotmail.com (S. Tan).

http://dx.doi.org/10.1016/j.eswa.2015.01.035
0957-4174/© 2015 Published by Elsevier Ltd.

Kotecha (2015, 2014) and Rather, Agarwal, and Sastry (2015), for instance, employed multiple or hybrid models involving SVR (Artificial Neural Regression), random forest and GA (Genetic Algorithms) to forecast stock market indicators.

Unfortunately, nonlinearity is not the only complication of an underlying stock market system which is also known to be timing varying and without a clear system boundary. Using an analogy, a stock market system shapes more like flowing water than standing mountain. Capturing nonlinearity in any fixed form alone is not sufficient to capture its nature. This is also why that time series forecast backed by nonlinear models or nonlinear model building techniques only works better for static off-line test data than for real stock market prediction situations.

In fact, these problems have long been realized and there have been constant efforts to find new solution strategies to address it. For instance, the concept of adaption has been introduced in prediction algorithms to tackle the time-varying nature of a stock market Hu, Feng, Zhang, Ngai, and Liu (2014) and more interesting idea of rule based adaptive modeling incorporating factors and nonlinear modeling structure has been studies more recently in handling nonlinearity, time-varying and boundary-less in one shot (Hu et al., 2014; Barak & Modarres, 2015). Interesting though, these ideas are still model-driven in nature, thus suffer from shortcomings, such as limitation and bias in factor selection, usually associated with model-driven techniques.

The preceding discussions have motivated us to consider reformulating the stock market prediction problem using a pure data-driven solution strategy aimed at finding a practical prediction solution directly from online data streams collected from any ongoing stock market. Here are the two key considerations underlying our new formulation. First, different from existing research with the purpose of quantitatively predicting stock market indicators, the prediction problem we consider is reduced to a narrower version of market trend prediction. This narrow problem appears to reduce significantly the depth of the prediction by limiting to qualitative trend changes only, thus a relatively easier but still practical prediction problem to solve. After all, if we are able to predict market trend change systematically and reliably, it will serve as a solid step for moving towards designing reliable schemes to predict actual market indices.

Second, as a stock market as a system is characteristically nonlinear, time-varying and boundary-less, the appropriate way to model it appears to use dominant set of factors and their inter-relationships represented in factor graph changing over time to capture its features essential for the required prediction tasks. Knowing the limitation of selecting factors based on expert knowledge, a more systematic and data-driven Bayesian Network model based on adaptation over a large factor set as developed by the authors of this paper (Wang, Wang, & Tan, 2008) is adopted for the model building.

In contrast to those function approximation inspired techniques like ANN with model structural elements often not bearing direct physical interpretations, Bayesian Network model is built on factors as its model structural elements. These factors carry direct meanings and allow domain specific interpretations. With Bayesian Network model structure, the model building becomes a process of establishing interconnection or conditional dependencies among a large set of meaningful factors amid a given online data stream collecting from a live stock market and the prediction is reduced to projecting key subset of factors that collectively have more significant impact on a stock market index at anytime of the prediction. This idea allows the seamless handling of nonlinearity, time-varying and boundary-less problems in one unified modeling framework. Indeed, the effectiveness of the Bayesian Network modeling combined with Ward clustering technique and K2 algorithm has been demonstrated in predicting stock prices and trend of stock index as in Zuo and Kita (2012b, 2012a), Zuo, Harada, Mizuno, and Kita (2012) and Eisuke, Harada, and Mizuno (2012), outperforming models such as Auto Regressive Moving Average model significantly in terms of accuracy.

In our work, to accommodate the complex time-varying nature our earlier technique has been expanded to compute a time series of emergent factor graphs, thus incorporating dynamics into the modeling and enabling the model to depict evolution of relationship among the factors. This concept is termed dynamical Bayesian factor graph and will be discussed in greater detail in the ensuing section.

In concise terms, the prediction problem we are trying to solve in this paper is simply formulated as predicting a market trend change using dynamical Bayesian factor graph as the model structure.

With the above problem formulation, we intend to develop a universal data driven approach to solve this market trend prediction problem in this paper. Our prediction solution consists of three main steps. First, select a set of macroeconomic factors as initial set of factors. Second, use dynamical Bayesian factor graph modeling algorithm developed by us to create a sequence of Bayesian factor graphs over a specified period of time. These Bayesian factor graphs will show the connecting topology of the dominant factors over the specified period of time. Third, identify drastic topological change over a few time adjacent graphs, and the points of changes in time will predict the changes in stock market trend in question.

The above solution has been applied to predicting the market trend changes of two major stock markets: the Shenzhen Stock market in China and the US stock market using the Shenzhen Stock Exchange Component Index and the US S&P 500 index, respectively, for about the same 11-year-period. Through empirical analysis, we have found that the topological structure of the factor graphs usually remains the same for consecutive months. Indeed, even for a long period of time, there are only fewer changes in the topological structure at fewer time points. Types of structural change are typically shown as: Some causal relationship may disappear or change direction, or new pair-wise link between two factors may show up, or existing links fade.

Our extensive computational results for both stock markets confirm that if the topological structure of the factor graphs remains unchanged, then the corresponding stock market trend will stay the same, whereas any significant topological structure change will signal a market trend change in a short distant future.

The rest of the paper is organized as follows. Section 2 describes the concept of dynamical Bayesian factor graph which is used as the model structure for market trend prediction. The actual prediction algorithm is also presented in this section. Section 3 conducts the empirical analysis for the market trend predictions using data from the Chinese and the US markets. Section 4 is the conclusion.

2. Dynamical Bayesian factor graph

We will start by briefly reviewing some basic concepts of Bayesian factor graph followed by expanding it to dynamical Bayesian factor graph to include time. This new dynamical graph with a set of factors adapted from an initial factor set is then used as the model structure for modeling a stock market for the purpose of trend prediction. The prediction algorithm is finally described on this new dynamical model structure.

2.1. Bayesian factor graphs

Being a subclass of Bayesian network (Pearl, 2000), Bayesian factor graph (Wang et al., 2008) is a probabilistic qualitative modeling structure that captures and represents relationships among a set of financial factors in a graphical format. The inter-factor relationships, by way of causality, relevance, or independence, are qualitatively represented by edges in a graph.

Mathematically, a Bayesian factor graph is a directed acyclic graph encoding the joint probability distribution for a set of factors in the form of random variables $X = \{X_1, \ldots, X_d\}$, where d is the number of factors in the set. This structure represents the qualitative information of a problem domain. The nodes in the graph represent random variables, and the absence of an edge between any two nodes represents conditional independency between the corresponding variables. If there is an edge from X_i to another node X_j, then X_i is a parent of X_j, and X_j is a child of X_i. The probability distributions between each node and its parents represent the quantitative information of a Bayesian factor graph.

The set of parent nodes of X_i is denoted by X_{G_i} and the whole graph structure is a vector $G = \{G_1, \ldots, G_d\}$. Given a graph structure G, the probability of X can be written as the product of the local distributions of each node X_i given its parents:

$$p(X|G) = p(X_1, \ldots, X_d|G) = \prod_{i=1}^{d} p(X_i|X_{G_i}). \tag{1}$$

For our problem, the graph structure G is not known. Introducing a prior $p(G)$ on the possible structures, we have

$$p(X) = \sum_G p(G)p(X|G) \tag{2}$$

By using the Bayes Rule, the posterior distribution of the graph structure becomes

$$p(G|X) = \frac{p(G)p(X|G)}{p(X)}. \tag{3}$$

$p(G \mid X)$ are termed scoring functions and can assume various specific forms. Based on discretized data, we use the Bayesian Dirichlet equivalence uniform scoring function (Heckerman, Geiger, & Chickering, 1995) for the marginal likelihood of the network $P(D \mid G, \alpha)$:

$$P(D|G,\alpha) = \prod_{i=1}^{d}\prod_{j=1}^{q_i} \frac{\Gamma\left(\frac{\alpha}{q_i}\right)}{\Gamma\left(\frac{\alpha}{q_i} + \sum_{k=1}^{r_i} N_{ijk}\right)} \prod_{k=1}^{r_i} \frac{\Gamma\left(\frac{\alpha}{q_i r_i} + N_{ijk}\right)}{\Gamma\left(\frac{\alpha}{q_i r_i}\right)}. \tag{4}$$

In (4), r_i is the number of nominal values of factor X_i, and q_i is the number of values of X_{G_i}. The $j_t h$ unique instantiation of G_i is denoted by ϕ_{ij}, and N_{ijk} is the number of records where factor X_i is instantiated as X_{ik} and X_{G_i} are instantiated as ϕ_{ij}. α is the equivalent sample size parameter which denotes the prior belief in the uniformity of the conditional distributions of the graph. This scoring function metric guides us to find the maximum posterior (MAP) structure for the set of factors under study.

Finding the optimal structure is to find a structure that maximizes the posterior probability

$$\widehat{G} \in \arg\max_G p(G|X). \tag{5}$$

which is the criterion used in Wang et al. (2008). There are various computational techniques in finding the structure G that maximize (5) (Friedman & Koller, 2003; Koivisto & Sood, 2004; Koivisto, 2006). The outcome of these techniques will deliver the optimal structure as the Bayesian factor graph model to the given set of factors.

An alternative approach to the scoring function described above is to compute the posterior probability of some local structural features, such as edges or node cliques or Markov-blankets (Friedman & Koller, 2003). It equals to the local summaries of the posterior distribution of the graph structures. More precisely, it equals

Table 1
Factor list.

Factor	Description
GDP	Gross domestic product
GDP_growth	GDP growth rate on a year-on-year basis
PPI	Producer price index
Inflation	Monthly inflation rate
Interest	Interest rate: 1-year treasury bill
Interest_growth	Interest rate growth rate on a year-on-year basis
CPI	Consumer price index
CCI	Consumer confidence index
Macro_index	Macro-economic climate index (for the Chinese market)
	Industrial production, capacity, and utilization (for US market)

$$p(f|X) = \sum_G p(G|X)f(G) \tag{6}$$

where f is the indicator of a structure feature. There are recent progresses in computing the posterior feature probabilities which are exact and efficient (Friedman & Koller, 2003; Silander, Kontkanen, & Myllymaki, 2007). It should be noted, however, that this technique will result in different graph structure G as compared to the scoring function based technique.

Note that both techniques are adaptable to on-line formulation over a given dataset. On-line adaptation also allows trimming to be incorporated in each step of the adaptation to eliminate edges that are below a pre-set threshold or not meeting a criterion, thus adapting the structure topology when new data are available. A direct benefit of this is that we can actually start with a very large initial factor set, the adaptation process through given data set will eliminate irrelevant or unimportant factors to deliver a right topological model structure with the dominant factors as a solution. This is the reason that this approach is called data-driven. In a data-driven situation like this, expert knowledge is no longer essential, although it can still be helpful at various stages of an adaptation to reduce irrelevant or unimportant factors.

2.2. Incorporating dynamics in Bayesian factor graph

Dynamics can easily be incorporated in the model structure of Bayesian factor graph discussed above.

Suppose we want to model the structural dynamics of d factors for a time period of (T_1, T_2) with a given data set D over the same time period: $D = \{X_{\tau i} : T_0 < T_1 \leqslant \tau \leqslant T_2,\ 0 < i \leqslant d\}$. For each t $(T_1 \leqslant t \leqslant T_2)$, there is one Bayesian factor graph G_t generated based on data set $D_t = \{X_{\tau i} : t_0 \leqslant \tau \leqslant t,\ 0 < i \leqslant d\}$. From D_t, the posterior probability is calculated for each possible edge $X_i \rightarrow X_j$ according to (6). Then the edges with a posterior probability greater than threshold p_0 are selected to constitute the Bayesian factor graph G_t. For a discrete time t_i of which $i = 1,2,\ldots,T$, and $T_1 < t_i < T_2$. This process will lead to a time series of Bayesian factor graphs: G_{ti} where $i = 1,2,\ldots,T$. These G_{ti} will be termed dynamical Bayesian factor graph that is seen as a dynamical model for an underlying system from the time period (T_1, T_2).

For the sake of our discussion on market trend prediction, it is important to be able to distinguish if two consecutive factor graphs in a time series of Bayesian factor graph is identical or not. To this end, the pair of Bayesian factor graphs G_{t-1} and G_t are compared with respect to their graphical structures. The number of new

Table 2
Time for the factors.

Market	T_0	T_1	T_2
Chinese	1997:01	1999:07	2008:06
US	1997:06	1999:12	2008:08

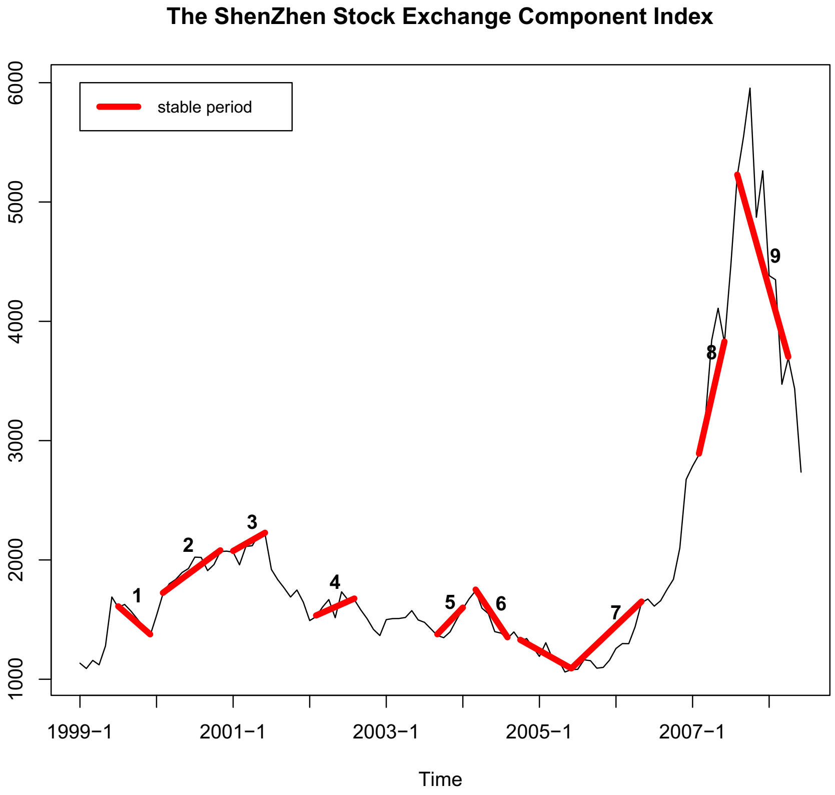

Fig. 1. Chinese market trend and structural stable periods.

edges and disappeared edges in G_t are counted. The two graphs are considered to be identical if there are no different edges in both graphs.

Another relevant issue is the development of efficient computational adaption algorithms to compute G_t. There have been some recent breakthroughs in this direction as well (Aliferis, Statnikov, Tsamardinos, Mani, & Koutsoukos, 2010a, 2010b). Needless to say that a prudent choice of initial set of factors will help to save a tremendous amount of computation. But the risk of bias in making an initial selection of factors, thus a significant error in prediction, sometimes outweighs the computational advantages gained.

2.3. Properties of dynamical Bayesian factor graph and prediction algorithm

G_t is seen as a new model structure that characterizes the underlying system with changing factor graph topology over time. It appears to be ideal in modeling systems that are nonlinear, time-varying and boundary-less, typical of financial market systems, for its feature of being able to capture factor relationships and evaluations of a complex system with respect to an observation target over a period of time.

With such a model, we can intuitively observe if there is any change in dominant factors of a system, which may actually be exactly the reason why an observed target is changed.

One important property of an underlying system resulting from this modeling approach has been summarized as follows:

Property 1. *If G_t is identical to G_{t-1}, the underlying system will stay on its course from $t-1$ to t, or more precisely, system variables under observation will not change their trends.*

At this time, we are only able to verify this property using extensive empirical analysis. It will be desirable to recast and prove it in a rigorous mathematical sense. But the interpretation of this property is obvious: If dominant factors of a system remain unchanged in their relationships, the behavior of the system will not change.

An interesting follow-on property is the following.

Property 2. *If G_t is not identical to G_{t-1}, the underlying system will change its course from $t-1$ to t, or more precisely, system variables under observation will change their trends. The degree of change will be determined by the degree of difference between G_t and G_{t-1}.*

Again, there is no rigorous proof of the above property at this time. We can only rely on empirical analysis to verify this property.

The above important property helps to link and trend change to the change of topological structure of G_t. This link is the basis for us to design the System Trend Prediction Algorithm below.

System Trend Prediction Algorithm:

Step 1: Select an initial set of factors $\{X_i\}$ for an underlying system.
Step 2: Compute its dynamical factor graph G_t using the data provided.
Step 3: Identify and mark the changes between G_t and G_{t-1} to predict the trend changes in the underlying system.

In the subsequent section, this algorithm will be applied to predict trend changes in two major stock market in the world: Shenzhen A-share market and US market using data for a period of over 10 years.

3. Computational experiments and analyses

The Stock Market data and the relevant macroeconomic data from both the China and US are used to implement the preceding trend prediction algorithm. The analysis will also be carried out to understand both the algorithm behavior and the results.

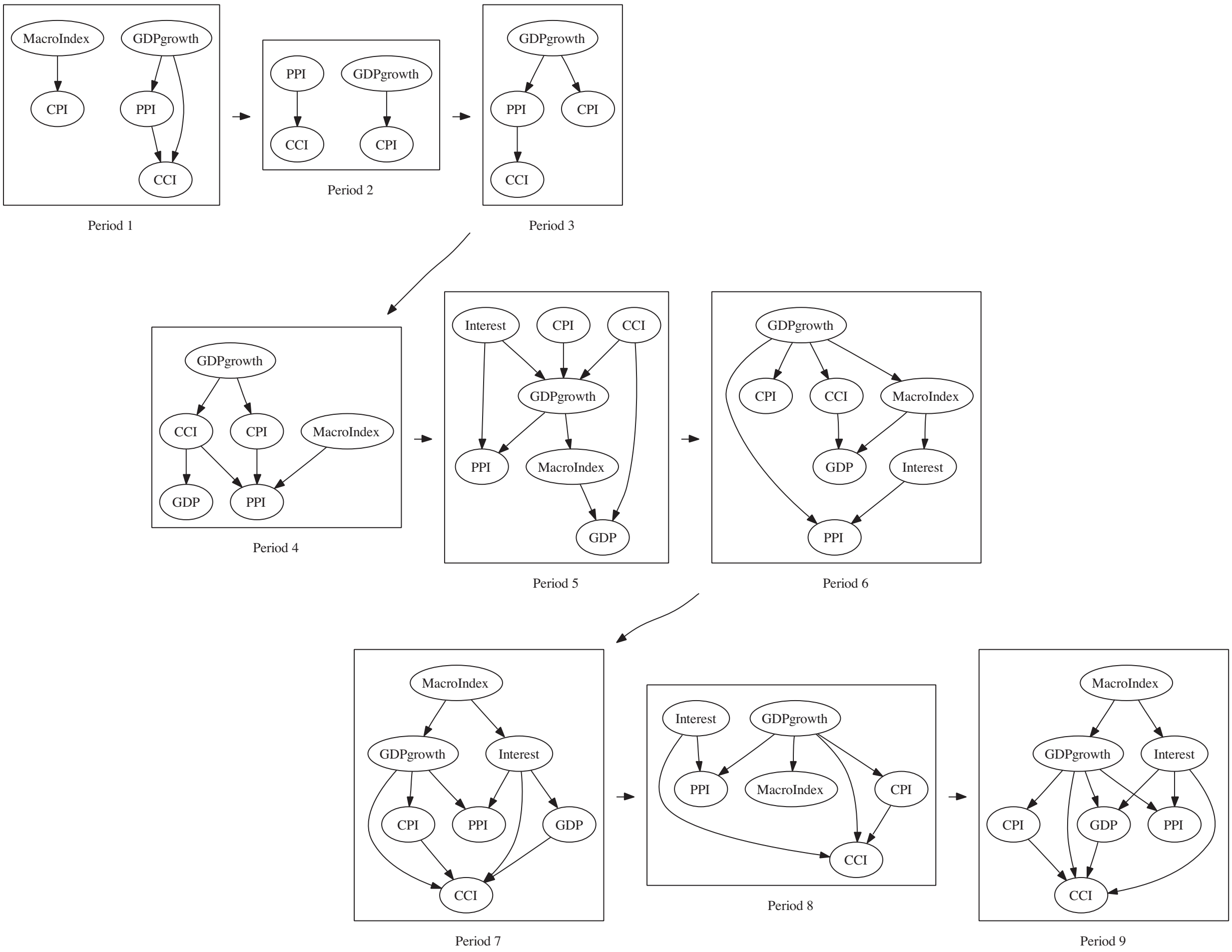

Fig. 2. Stable structural evolution for the Chinese market.

3.1. Data preparation

To start, we need to select the initial set of factors. Then the data can be collected for these factors over a preset time frame and interval for the prediction algorithm to use. A total of 9 macroeconomic factors have been selected as the underlying dominant factors for both the Chinese stock market and the US stock market as listed in Table 1. These factor cover different aspects of the economy: consumption, production, inflation, and interest, etc. and are known to have direct impact on stock markets for a given economy. If there is no any prior knowledge to help with the initial selection of relevant factors, one could start with a very large initial factor set and let the adaptation algorithm to use the data to build its own Bayesian factor graph through convergence. This latter approach will require much larger data set and more computational time to arrive at smaller and manageable size of a factor graph. The time periods for the factors in the two markets are in Table 2 with monthly intervals.

The index data and all the other relevant data are collected for the stated durations in monthly intervals for the Chinese Shenzhen A-Share market indexed by the Shenzhen Stock Exchange Component Index and the US stock market indexed by S&P500. The data sources used for the factor data and index data are from finance.sina.com.cn. For convenience, we form two datasets consisting of factor data and index data, one for the Chinese market and the other for the US market.

As shown in Table 2, for each dataset we use the first 30 months $[T_0, T_1]$ to be the out-of-sample period, i.e., $T_1 = T_0 + 30$ months, while $[T_1, T_2]$ is the in-sample-period. Samples of variables are first discretized into nominal values before they are fed into the adaptive algorithm to produce the Bayesian factor graph.

There is one factor graph learned for each month. For month $t \in [T_1, T_2]$, we use the data during $[t - 30, t - 1]$ to build the graph. The posterior probabilities of all the edges are calculated first. We then select the edges with a posterior probability greater than 0.5 as a preset threshold. Therefore, for consecutive months when one edges posterior probability goes above or falls below 0.5, there will be a topological structural change.

3.2. Computational results and analysis for the Chinese stock market

For the Chinese market for the duration between 1999:07 and 2008:06, there is one Bayesian factor graph learned for each month. To compare the topological structural evolutions and changes of Bayesian factor graphs in relationship to the Shenzhen market, we use the Shenzhen Stock Exchange Component Index as backdrop and use redline to mark a period where there is no topological structural change within the same period as in Fig. 1. Such a period will be called stable period.

As shown in Fig. 1, there are 9 stable periods marked in the bold line segment:

Period 1: 1999:07–1999:12, Period 2: 2000:02–2000:11, Period 3: 2000:12–2001:06,
Period 4: 2002:02–2002:08, Period 5: 2003:09–2004:01, Period 6: 2004:03–2004:08,

Table 3
The edge posterior probabilities in each graph for the Chinese market.

Graph	Time	Edge	Posterior probability
Graph 1	1999:10	GDPgrowth → PPI	0.71
		Macro Index → CPI	0.66
		GDPgrowth → CCI	0.56
		PPI → CCI	0.55
Graph 2	2000:07	GDPgrowth → CPI	0.88
		PPI → CCI	0.84
Graph 3	2001:03	PPI → CCI	0.90
		GDPgrowth → CPI	0.85
		GDPgrowth → PPI	0.62
Graph 4	2002:05	MacroIndex → PPI	0.89
		CPI → PPI	0.88
		GDPgrowth → CPI	0.85
		GDPgrowth → CCI	0.74
		CCI → PPI	0.71
		CCI → GDP	0.63
Graph 5	2003:12	GDPgrowth → MacroIndex	0.97
		CCI → GDPgrowth	0.96
		Interest → GDPgrowth	0.94
		GDPgrowth → PPI	0.93
		CCI → GDP	0.92
		CPI → GDPgrowth	0.91
		Interest → PPI	0.85
		MacroIndex → GDP	0.67
Graph 6	2004:06	GDPgrowth → PPI	0.95
		GDPgrowth → CPI	0.92
		Interest → PPI	0.82
		GDPgrowth → CCI	0.79
		CCI → GDP	0.76
		MacroIndex → Interest	0.72
		MacroIndex → GDP	0.70
		GDPgrowth → MacroIndex	0.65
Graph 7	2005:07	GDPgrowth → CCI	0.98
		CPI → CCI	0.97
		Interest → CCI	0.97
		GDPgrowth → PPI	0.94
		GDPgrowth → CPI	0.93
		Interest → PPI	0.84
		GDP → CCI	0.79
		Interest → GDP	0.77
		MacroIndex → Interest	0.68
		MacroIndex → GDPgrowth	0.60
Graph 8	2007:02	GDPgrowth → PPI	0.97
		Interest → PPI	0.91
		GDPgrowth → CCI	0.87
		GDPgrowth → CPI	0.83
		CPI → CCI	0.73
		GDPgrowth → MacroIndex	0.70
		Interest → CCI	0.51
Graph 9	2007:12	GDPgrowth → CCI	1.00
		CPI → CCI	1.00
		Interest → CCI	1.00
		GDP → CCI	1.00
		GDPgrowth → PPI	0.95
		Interest → GDP	0.94
		GDPgrowth → CPI	0.93
		Interest → PPI	0.88
		GDPgrowth → GDP	0.77
		MacroIndex → Interest	0.68
		MacroIndex → GDPgrowth	0.56

Period 7: 2004:10–2006:05, Period 8: 2007:02–2007:06, Period 9: 2007:07–2008:04.

Corresponding to within these 9 stable periods, the stock market can be observed to stay on with its trend, i.e, generally the market index keeps rising or falling within any such a stable Period. However, between any of the two stable Periods, there is single or multiple Bayesian factor graph structural changes as captured by our algorithm. Observe that the length of the structural changing periods, called skittish periods, vary from 1 to about 6 months. One example of a relatively long skittish period is 2003:03–2003:07, during which the structures keep changing from month to month for four consecutive months, whereas for 2004:02–2004:03 and 2007:07, the skittish periods only last for one or two months. It is interesting to note that the beginning or ending points of either a stable period or a skittish period marks the turning points of the general market trends and are termed trend prediction points. Note that our current algorithm can only determine trend changes, but not directions of the changes.

One of the most important observations, as exhibited by our computational results, is that the trend prediction points always precede the actual turning points of the market index by a lead margin of a few months. For example, before the sharp drop in Period 9 in Fig. 1, the Bayesian factor graph structural change occurs in 2007:07, then remains invariant during 2007:07–2008:04. There is a three-month advance of the Bayesian factor structure change before the historical market peak at 6124.04 in 2007:10.

Fig. 2 shows all 9 computed stable Bayesian factor graphs corresponding to the 9 Periods marked in Fig. 1. In these stable graphs, the factors GDP growth rate and Macro Index are always at higher levels in the graph hierarchy, while Consumer Confidence Index (CCI) is at the bottom for most of the times. The disappearing or emerging edges are interpreted as the driving force of the market change. A drastic change in the structure is usually a significant sign of a forthcoming market change. For example, comparing the two graphs for Period 8 and Period 9 with 11 and 7 edges respectively, and the graph for Period is different with 4 new edges emerging and 1 edge reversing direction. Accordingly, the market index totally changes the direction after Period 8 and moving to Period 9.

The edge posterior probabilities in all 9 factors graphs are shown in Table 3. The posterior probability is interpreted as measure of the relationship between two factors linked by this edge. It should be noted that even in the 9 stable periods, the posterior probabilities of the same edge are different in different months. For simplicity, Table 3 only shows the result of one month in the middle of the period for illustration. From Table 3, it can be seen that there are some strong relationships with very high posterior probabilities in the graphs (greater than 0.9) and the edge posterior probabilities also change in different graphs along with time.

3.3. Computational result and analysis for the US stock market

Following the above conventions, the stable periods for the US market are computed and shown in Fig. 3 against the backdrop of the S&P 500 index. It can be seen that for the US market the observation that a stable graph corresponds to a stable general market trend still stands. There are 6 stable periods with stable Bayesian factor graphs in Fig. 3 marked as bold line segments:

Period 1: 2002:01–2002:07, Period 2: 2003:03–2004:03, Period 3: 2004:06–2004:10,
Period 4: 2005:10–2006:04, Period 5: 2006:06–2007:03, Period 6: 2008:02–2008:06.

Similar to the Chinese market, the lengths of the changing period vary, from 2 months to 1 year. However, both the stable periods and the skittish periods appear fundamentally different in time for the two different markets. For example, the US market the Bayesian factor graphs are stable in Period 5 (2006:06–2007:03), whereas for the Chinese market the factor graphs are almost always changing during the same time.

Like the results of the Chinese market, the starting and ending times of stable periods always precede the real turning points of

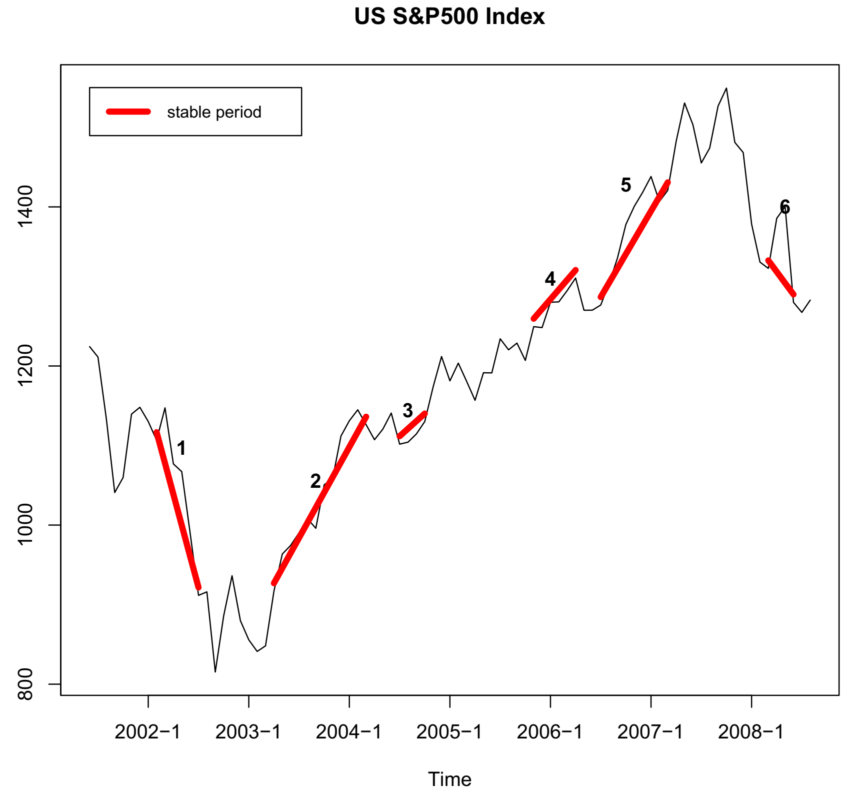

Fig. 3. US market trend and structural stable periods.

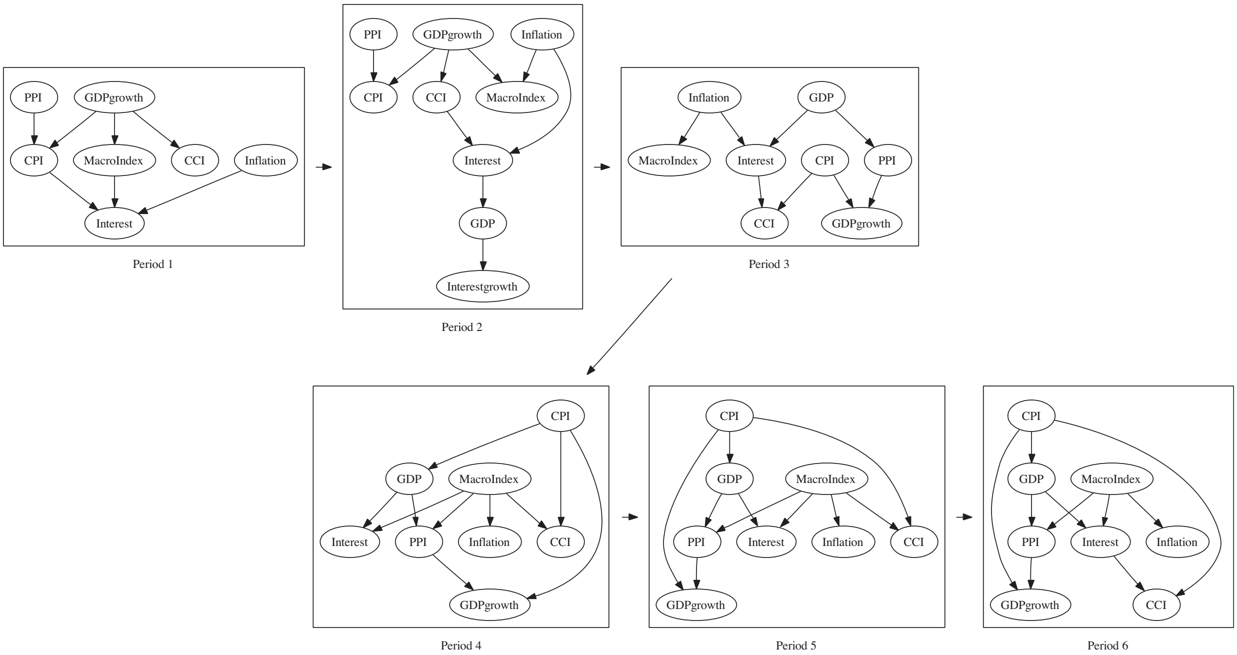

Fig. 4. Stable structural evolution for the US market.

the market trend. This is particularly obvious for the market falling periods: Period 1 and 6. There are about 2 months advance before the market reaches its local peak and starts falling down. This kind of advanced structural change is therefore suitable potential candidate as advance warning for investors about a possible trend change in the market.

The stable factor graph structures for the 6 stable periods are shown in Fig. 4. It appears that the structural differences of the first 3 periods are more significant than the last 3 periods. For Periods 4, 5 and 6, the graphs share similar general topology with local variations only.

Comparing the two markets, it can also be seen that the resulting factor relationships for the US market are significantly different from those for the Chinese market. Even for Period 5 (2003:09–2004:01) for the Chinese market and Period 2 (2003:03–2004:03) for the US market with overlapping time ranges, the differences

Table 4
The edge posterior probabilities in each graph for the US market.

Graph	Period	Edge	Posterior probability
Graph 1		Inflation → Interest	0.97
		GDPgrowth → MacroIndex	0.93
		GDPgrowth → CCI	0.70
		GDPgrowth → CPI	0.69
		PPI → CPI	0.64
		CPI → Interest	0.55
		MacroIndex → Interest	0.53
Graph 2		Inflation → Interest	0.97
		GDPgrowth → MacroIndex	0.94
		Inflation → MacroIndex	0.88
		Interest → GDP	0.85
		CCI → Interest	0.85
		GDP → Interestgrowth	0.83
		GDPgrowth → CPI	0.78
		PPI → CPI	0.63
		GDPgrowth → CCI	0.55
Graph 3		CPI → GDPgrowth	0.87
		PPI → GDPgrowth	0.85
		Interest → CCI	0.70
		GDP → Interest	0.68
		CPI → CCI	0.65
		Inflation → MacroIndex	0.64
		GDP → PPI	0.59
		Inflation → Interest	0.55
Graph 4		GDP → Interest	1.00
		CPI → GDPgrowth	1.00
		MacroIndex → Interest	1.00
		PPI → GDPgrowth	0.99
		MacroIndex → Inflation	0.96
		GDP → PPI	0.96
		CPI → GDP	0.77
		CPI → CCI	0.74
		MacroIndex → CCI	0.71
		MacroIndex → PPI	0.66
Graph 5		CPI → GDPgrowth	1.00
		MacroIndex → Interest	1.00
		GDP → Interest	1.00
		PPI → GDPgrowth	0.99
		MacroIndex → Inflation	0.97
		GDP → PPI	0.96
		CPI → CCI	0.88
		CPI → GDP	0.75
		MacroIndex → PPI	0.67
		MacroIndex → CCI	0.60
Graph 6		CPI → CCI	0.99
		MacroIndex → Interest	0.98
		Interest → CCI	0.98
		MacroIndex → Inflation	0.97
		GDP → Interest	0.96
		CPI → GDPgrowth	0.71
		PPI → GDPgrowth	0.71
		GDP → PPI	0.68
		CPI → GDP	0.62
		MacroIndex → PPI	0.50

of the two structures are still very significant. This significant difference, from the Bayesian factor graph modeling point of view, confirms the general belief that the Chinese and US markets are driven with very distinct mechanisms.

Table 4 shows the posterior probabilities of the edges in all 6 factor graphs for the US market from our computation. Again the result of the middle month of each period is shown for the sake of simplicity. Compared with the results for the Chinese market, there are more edges in each stable graph and more reliable edges with very high posterior probabilities (close to 1) for the US market.

Apart from the Shenzhen Stock Exchange Component Index and S&P500 index, we also have performed the similar computation analysis for other markets such as the Shanghai Stock Exchange and NASDAQ drawing similar results.

4. Conclusions

We have developed a new method in predicting market trends using dynamical Bayesian factor graphs and demonstrated its predictive power through computational analysis based on both Chinese and US stock markets. Being qualitative in nature, this method appears to be able to capture the change in market trends effectively. Comparing with the prediction techniques that use fixed nonlinear mathematical structures, the resulting Bayesian factor graph of our method at different time instance appears to be able to provide the insight as to the factors dominating a market at that particular instance and a change in factors graph topology through a period of time depict how a market is undergoing change in its inner driving force, which will eventually show up as the change in market trend. Our research sheds some new light on forecasting market trend, and hopefully could aid investors to more effectively design trading strategies.

Our method is still at an earlier stage and requires more extensive work to make it well-understood and a more reliable practical tool for analysts as well as investors. In particular, the theoretical basis of the structural change in dynamical Bayesian factor graph needs to be further explored for a better understanding this modeling mechanism. It remains open for such a theoretical framework to aid the analysis. It appears that one possible theoretical path is to use certain concepts in Graph Dynamical System in representing and understanding structural changes of dynamical graphs, which we are currently pursing. As discussed, one limitation with our current method is that we are unable to ascertain the direction of a trend change even if we know there is a trend change coming up with our computation. We hope to be able to ascertain a direction of change through a better theoretical understanding of the underlying process.

The intensity in computation of our modeling method in the presence of a very large factor set is another issue that needs to be addressed. For its data-driven nature and without expert guidance, a typical factor set can be as large as a few hundred initial factors. We are working on creating a dominate subset selection algorithm to turn the problem of eliminating the number of irrelevant factors to an optimization problem of locating a dominant subset using an appropriate utility target.

It is our belief the notion of dynamical Bayesian factor graph is a potentially powerful modeling structure for very complex systems that are nonlinear, time-varying and boundary-less, hence may be well-suited for modeling other financial market dynamics for prediction and other purposes, such as the market of foreign currency exchange and commodities markets.

References

Aliferis, C., Statnikov, A., Tsamardinos, I., Mani, S., & Koutsoukos, X. (2010a). Local causal and Markov blanket induction for causal discovery and feature selection for classification. Part II: Analysis and extensions. *Journal of Machine Learning Research, 11*, 171–234.

Aliferis, C., Statnikov, A., Tsamardinos, I., Mani, S., & Koutsoukos, X. (2010b). Local causal and Markov blanket induction for causal discovery and feature selection for classification. Part II: Analysis and extensions. *Journal of Machine Learning Research, 11*, 235–284.

Barak, S., & Modarres, M. (2015). Developing an approach to evaluate stocks by forecasting effective features with data mining methods. *Expert Systems with Applications, 42*(3), 1325–1339.

de Oliveira, F. A., Nobre, C. N., & Zrate, L. E. (2013). Applying artificial neural networks to prediction of stock price and improvement of the directional prediction indexCCase study of PETR4, Petrobras, Brazil. *Expert Systems with Applications, 40*(18), 7596–7606.

Eisuke, K., Harada, M., & Mizuno, T. (2012). Application of Bayesian network to stock price prediction. *Artificial Intelligence Research, 1*(2), 171–184.

Fama, E., & French, K. (1996). Multifactor explanations of asset pricing anomalies. *Journal of Finance, 51*, 55–84.

Friedman, N., & Koller, D. (2003). Being bayesian about network structure: A bayesian approach to structure discovery in bayesian networks. *Machine Learning, 50*, 95–126.

Heckerman, D., Geiger, D., & Chickering, D. M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning, 20*(3), 197–243.

Huang, W., Nakamori, Y., & Wang, S. (2005). Forecasting stock market movement direction with support vector machine. *Computers and Operations Research, 32*(10), 2513–2522.

Huang, C., & Tsai, C. (2009). A hybrid SOFM-SVR with a filter-based feature selection for stock market forecasting. *Expert Systems with Applications, 36*(2), 1529–1539.

Hu, Y., Feng, B., Zhang, X., Ngai, E., & Liu, M. (2014). Stock trading rule discovery with an evolutionary trend following model. *Expert Systems with Applications, 42*(1), 212–222.

Kim, K. (2003). Financial time series forecasting using support vector machines. *Neurocomputing, 55*(1), 307–319.

Kim, K. (2003). Financial time series forecasting using support vector machines. *Neurocomputing, 55*(1), 307–319.

Koivisto, M. (2006). In *Proceedings of the 22nd conference on uncertainty in artificial intelligence* (pp. 241–248).

Koivisto, M., & Sood, K. (2004). Exact Bayesian structure discovery in Bayesian networks. *Journal of Machine Learning Research, 5*, 549–573.

Lettau, M., & Ludvigson, S. (2001). Resurrecting the (C)CAPM: A cross-sectional test when risk premia are time-varying. *Journal of Political Economy, 109*(6), 1238–1287.

Lettau, M., & Ludvigson, S. (2005). Expected returns and expected dividend growth. *Journal of Financial Economics, 76*, 583–626.

Naranjo, A., Nimalendran, M., & Ryngaert, M. (1998). Stock returns, dividend yields and taxes. *Journal of Finance, 53*, 2029–2057.

Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2014). Predicting stock market index using fusion of machine learning techniques. *Expert Systems with Applications, 42*(4), 2162–2172.

Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert Systems with Applications, 42*(1), 259–268.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge: Cambridge University Press.

Rather, A. M., Agarwal, A., & Sastry, V. (2015). Recurrent neural network and a hybrid model for prediction of stock returns. *Expert Systems with Applications, 42*(6), 3234–3241.

Santos, T., & Veronesi, P. (2006). Labor income and predictable stock returns. *Review of Financial Studies, 19*(1), 1–44.

Silander, T., Kontkanen, P., & Myllymaki, P. (2007). In *Proceedings of the 23rd conference on uncertainty in artificial intelligence* (pp. 360–367).

Tay, F., & Cao, L. (2001). Application of support vector machines in financial time series forecasting. *Omega, 29*(4), 309–317.

Ticknor, J. L. (2013). A Bayesian regularized artificial neural network for stock market forecasting. *Expert Systems with Applications, 40*(14), 5501–5506.

Wang, Z., Wang, L., & Tan, S. (2008). Emergent and spontaneous computation of factor relationships from a large factor set. *Journal of Economic Dynamics and Control, 32*(12), 3939–3959.

Żbikowski, K. (2014). Using volume weighted support vector machines with walk forward testing and feature selection for the purpose of creating stock trading strategy. *Expert Systems with Applications, 42*(4), 1797–1805.

Zuo, Y., Harada, M., Mizuno, T., & Kita, E. (2012). Bayesian network based prediction algorithm of stock price return. In *Intelligent decision technologies* (pp. 397–406). Springer.

Zuo, Y., & Kita, E. (2012a). Stock price forecast using Bayesian network. *Expert Systems with Applications, 39*(8), 6729–6737.

Zuo, Y., & Kita, E. (2012b). Up/down analysis of stock index by using Bayesian network. *Engineering Management Research, 1*(2), 46–52.