4. STOCK INDEX PATTERN DISCOVERY VIA TOEPLITZ INVERSE COVARIANCE-BASED CLUSTERING

Hongbing OUYANG[1]
Xiaolu WEI[2]
Qiufeng WU[3]

Abstract

In this study, we attempt to discover repeated patterns of stock index through analysis of multivariate time series. Our motivation is based on the notion that financial planning guided by pattern discovery of stock index may be more effective. A two-stage architecture constructed by combining Toeplitz Inverse Covariance-Based Clustering (TICC) with betweenness centrality is applied for pattern discovery of stock index. In the first stage, TICC is used to discover repeated patterns of stock index's multivariate time series. Then, in the second stage, betweenness centrality scores that reveal the relative "importance" of each influencing factor are plotted by Markov random field (MRF) which is derived from the first experiment. The Hangseng Stock Index and five influencing factors are used in the experiment. Empirical results show that six kinds of time-invariant patterns with flexible and long-term time periods are discovered in Hangseng Stock Index, and that different influencing factors have different betweenness scores in each cluster. This empirical research provides new ideas of stock index prediction and portfolio construction for scholars and investors. Moreover, the long-period repeated patterns that are discovered in this paper increase the possibility to reduce transaction cost for portfolio construction which then, maybe more applicable to the real financial markets.

Keywords: pattern discovery, Toeplitz Inverse Covariance-Based Clustering (TICC), multivariate time series, stock index, influencing factors, betweenness centrality scores

JEL Classification: C38, C63, G17

[1] *School of Economics, Huazhong University of Science and Technology, Hubei Province, P.R. China.*
[2] *Corresponding author. Business School, Hubei University, Hubei Province, P.R. China. Email: 383212186@qq.com.*
[3] *School of Science, Northeast Agricultural University, Heilongjiang Province, P.R. China.*

1. Introduction

Stock index can accurately reflect current market trends in a timely manner. Therefore, Stock index prediction is one of the most important subjects in financial time series forecasting (Moghaddam *et al.*, 2016). However, stock index characteristics, noisy and non-stationary, make prediction face challenges. 'Noisy' implies that there is insufficient information for investors to observe past behaviors of stock index. 'Non-stationary' means that stock index may change dramatically in different periods. These characteristics lead to poor stock index prediction results as predicted by traditional econometric models such as linear model, Auto-Regressive Integrated Moving Average (ARIMA) and Vector Auto-Regression (VAR) (Shen, 2016; Bouri, 2018). The aforementioned methods belong to short-term predictions in time series, which are seriously affected by 'Noisy' and 'non-stationary'. However, if stock index prediction only focuses on forecasting the trend over a certain period, effects of 'noisy' and 'non-stationary' on the prediction results will be weakened. One method of stock index trend prediction over a certain period is to decompose stock index over long-period into many stock index fragments over short-period, and then stock index fragments over short-period are classified into some pattern in pattern sets including "head and shoulders", "W" and "spine", etc.. The core of the above-mentioned process is to find some inherent patterns in the existing stock index sequences, which is referred to as "pattern discovery" (Chen, 2016). According to the discovered patterns, we can predict the future trend of stock index over a certain period, and take investment actions more appropriately.

Recently, more and more scholars analyze repeated patterns of time series through machine learning techniques. Unlike traditional econometric models, machine learning techniques are based on fewer assumptions, which allow for models to better fit dynamic processes in time series (Chou and Ego, 2016). Scholars can explore complex structures in time series and investigate the relationships between variables more effectively through machine learning techniques. Therefore, machine learning techniques maybe more appropriate to discover the repeated patterns of stock index.

However, machine learning techniques are seldom used in the pattern discovery of stock index. Based on the notion that there are interactions among different factors in stock index, and inherent patterns in stock index subsequences are time invariant (Sezer *et al*., 2017), the Toeplitx Inverse Covariance-Based Clustering (TICC) (Hallac *et al*., 2017) technique that considers factor interaction and time-invariance in stock index might be more suitable for pattern discovery of stock index. Therefore, in this study, we attempt to discover repeated patterns of stock index through the TICC technique. Specifically, pattern discovery of stock index through TICC consists of two phases. In the first phase, we cluster influencing factors of stock index based on their interaction. In the second phase, we map the cluster results to stock index. Through these two phases, we could discover repeated patterns of stock index more comprehensively.

The rest of this paper is organized as follows. Section 2 summarizes the problems found in current research, and puts forward the contributions of this paper. Section 3 describes the mathematical model on pattern discovery of stock index and detailed introduction to TICC method. Section 4 describes the experimental preliminaries, which contain experimental data and selection of influencing factors. Section 5 presents the application steps of TICC and the experimental results on Hangseng stock index. Finally, the conclusion is drawn in section 6.

2. Literature Review

Since the first study of Malkiel and Fama (1970), the predictability of financial markets has been the focus of scholars' in-depth study. In Malkiel and Fama's research (1970), they proposed the efficient market hypothesis and denied the predictability of financial market. However, anomalies appeared subsequently in financial markets (Avramov and Chordia, 2006). In order to explain these anomalies, scholars proposed behavioral finance and adaptive market hypotheses (Tversky and Kahneman, 1979; Lo, 2004). Based on these theories, many scholars conducted empirical research and confirmed the predictability of financial markets and stock index (Bannigidadmath and Narayan, 2016; Mclean and Pontiff, 2016). However, due to the high noise and non-stationary characteristics of stock index, scholars face challenges to predict stock index through traditional econometric methods. This is partly due to the strict assumptions in traditional econometric methods, such as linearity, stationarity, normal distribution, etc. (Shen *et al*., 2016). These assumptions are inconsistent with real financial market. Therefore, scholars may fail to predict stock index accurately through econometric methods.

Pattern discovery and prediction of stock index through machine learning techniques may solve the above problems. Pattern discovery described here is the act of finding recurrent patterns that occur in time series (Min *et al*., 2020). In the field of pattern discovery, there are various approaches, such as probability dense function (Parzen, 1962) and localized radial basis functions (Fu *et al*., 2001). These techniques investigate a non-parametric description of data structure through locally adjusted tuned basis functions. In other words, any recurrent pattern dominated by a functional relationship can be discovered through unlimited numbers of hidden units and samples. The studies related to pattern discovery in time series analysis mainly have three categories. One is to identify statistically significant events in time series (Wieldraaijer *et al*., 2018). The second one is to seek inherent structure in time series (Baumeister *et al*., 2020). The third one is to find pattern rules from time series data (Guo *et al*., 2017).

Pattern discovery techniques have been widely used in various fields such as visualization and telecommunication (Jhaveri *et al*., 2018). However, few studies are associated with finding repeated patterns of stock index. These studies mainly focus on three areas. The first one is to find repeated patterns of stock index with fixed time period based on clustering techniques (Cen and Qin, 2016). The second one is to find repeated patterns of stock index through technical analysis, which analyze stock index prices and volumes in the history (Schumaker *et al*., 2017). The third one is to discover special event patterns through some templates, such as stock market bulletin (Nisar and Yeung, 2018).

In addition, we find three problems in the literature on stock index pattern discovery. The first problem is that existing financial studies mainly focus on pattern discovery of univariate time series without considering the influence of other factors (Arévalo *et al*., 2017). The second problem is that some studies on multivariate time series usually discover repeated patterns with fixed time period through distance-based measurements. These measurements have risk of overfitting, and may produce unsatisfactory results in certain situations (Cen and Qin, 2016). The third problem, which is also the most important one, is that these studies cannot interpretate the results of pattern discovery based on subsequence clustering, especially when the data is highly-dimensional (Hallac *et al*., 2017).

In order to solve the problems mentioned in previous literature, this paper intends to discover repeated patterns of stock index through the TICC algorithm. The contributions of this paper are as follows.

- Assume that each cluster that defines a pattern is time-invariant over a window of size w, which could be accomplished through Markov Random Field (MRF) in TICC. Here, MRF is a kind of correlation network with multiple layers, which represents time-invariant correlation structure for each cluster through a constrainted inverse covariance estimation problem and a block Toeplitz structure (Friedman *et al*., 2008).
- Discover repeated patterns of stock index with full consideration of different influencing factors and the interdependency among these factors.
- Discover repeated patterns with flexible and long-term time period through cluster assignment of TICC technique, which may help to improve generalization ability of the model, and reduce transaction cost for portfolios in subsequent research on pattern prediction and portfolio construction.
- Interpretate the results of repeated patterns more efficiently through MRF, betweenness centrality score and related graphic theories in technical analysis.

3. Stock Index Pattern Discovery Based on TICC

3.1 Mathematical Model of Stock Index Pattern Discovery

This paper focuses on discovering stock index pattern and analyzing the relative importance of each influencing factor. Dataset on stock index and influencing factors mainly has two characteristics: correlation and time-invariance. Correlation could be solved by constructing MRF and Toeplitz matrix for each cluster. Time-invariance could be solved by introducing parameter β, which incentivizes nearby subsequences to be assigned to the one cluster. The detailed mathematical model on stock index pattern discovery is described as follows.

Given datasets $\{\mathrm{X_{org_t}}\}_{\mathrm{t=1}}^{\mathrm{n}}$, wherein $\mathrm{X_{org_t}} = (\mathrm{y_t}; \mathrm{x_{t1}}, \mathrm{x_{t2}}, \cdots, \mathrm{x_{tm}})^{\mathrm{T}}$, the process of pattern discovery is a clustering problem. Instead of looking at x_t only, a short subsequence of size w (w<T, wherein T is the length of time series) is clustered. This subsequence includes observations going from time t-w+1 to t, which is called $\mathrm{X_t}$. We call this new sequence X, going from $\mathrm{X_1}$ to $\mathrm{X_t}$. Given new datasets $\{\mathrm{X_t}\}_{\mathrm{t=1}}^{\mathrm{n}}$, TICC method is applied to get cluster results which is denoted by $\{\mathrm{C_i}\}_{\mathrm{i=1}}^{\mathrm{a}}$, and Makov random field $\{\mathrm{MRF_i}\}_{\mathrm{i=1}}^{\mathrm{a}}$ of each cluster. Based on the Makov random field $\{\mathrm{MRF_i}\}_{\mathrm{i=1}}^{\mathrm{a}}$ of each cluster, we could analyze the relationship between stock index and other influencing factors. Moreover, we could investigate the relative importance of each influencing factor through betweeness centrality score which is denoted by $(\mathrm{RI_{im}})_{\mathrm{M\times a}}$, which might be helpful in the future research on stock index pattern prediction.

3.2 Toeplitz Inverse Covariance-Based Clustering Method of Stock Index Pattern Discovery

Clustering is widely used for pattern discovery. In the clustering process of pattern discovery, three problems must be considered. The first one is the length of different patterns. Many studies prefer pattern templates of fixed length for representing repeated patterns rather than flexible length which is used by our method. The second one is the risks of overfitting which usually yield unsatisfactory outcomes. The last problem is the efficiency of discovery process. With the increase of data size, the time needed to discover repeated patterns

increases remarkably. Therefore, an algorithm to reduce the total number of data points and the operation time of pattern discovery is indispensable. These three problems are closely related to the clustering process of pattern discovery.

TICC was proposed by Hallac *et al*. in 2017, which was runner up in the research track of Knowledge Discovery and Data Mining (KDD). Compared to other clustering methods, TICC is the first method to cluster time series based on graphical dependency structure, which could solve the above-mentioned problems. In this section, we describes the details of TICC algorithm applied in this paper.

TICC is a clustering technique that cluster short subsequences of size w (w<T, wherein T is the length of time series), going from time t-w+1 to t. In this paper, window size w and cluster number k of stock index are chosen to be 1 and 3, respectively. Each cluster is defined based on a Gaussian inverse covariance $\Theta_i \in R^{n\times n}$ (i=1, 2, 3), which describes the structural representation of cluster i (Koller and Friedman, 2009). The objective of TICC method applied in this paper is to solve the covariance of each cluster $\Theta \in \{\Theta_1, \Theta_2, \Theta_3\}$ and the assignment results $P \in \{P_1, P_2, P_3\}$, where $P \subset \{1,2,\cdots,T\}$. The objective is achieved through two steps: cluster assignment and Toeplitz graphical lasso. In the Cluster assignment, TICC solves the subproblem through a dynamic programming algorithm. In the Toeplitz graphical lasso, TICC updates the cluster parameters based on the alternating direction method of multipliers (ADMM). This combinatorial algorithm is equivalent to expectation maximization (EM). Based on the assignment results, we could discover recurrent clusters of financial multivariate time series and make appropriate actions. The optimization problem in this paper is written as follow:

$$\operatorname{argmin}_{\Theta\in\Gamma,P} \sum_{i=1}^{3} \left[\overbrace{\|\lambda \circ \Theta_i\|_1}^{\text{sparsity}} + \sum_{X_i\in P_i} \left(\overbrace{-\lambda\lambda(X_t, \Theta_i)}^{\text{log likelihood}} + \overbrace{\beta I\{X_{t-1} \notin P_i\}}^{\text{temporal consistency}} \right) \right].$$

Here, Γ is Toeplitz matrices, λ ($\lambda \in R^{n\times n}$) is a parameter that determines the sparsity level in the MRFs. In other words, we could minimize the negative log likelihood and make sure Θ_i (i=1, 2, 3) is sparse based on regularization parameter λ. β is another parameter that incentivize neighboring points to be assigned to one cluster, the neighboring subsequences are more likely to belong to one cluster as β gets larger. In this paper, λ and β are chosen to be 11e-2 and 600, respectively.X_t is input matrix which include stock index and 5 influencing factors. We will give details of influencing factors in Section 3. $\|\lambda \circ \Theta_i\|_1$ (i=1, 2, 3) is a ℓ_1-norm penalty to encourage a sparse Θ_i (i=1, 2, 3) and prevent overfitting (Steffen, 1996). $-\lambda\lambda(X_t, \Theta_i)$(i=1, 2, 3) is the negative log likelihood of X_t comes from cluster i (i=1, 2, 3). $I\{X_{t-1} \notin P_i\}$ (i=1, 2, 3) is a function that analyze whether adjacent subsequences are assigned to one cluster.

3.2.1 Cluster Assignment

Given the inverse covariances Θ_i (i=1, 2, 3) and one of regularization parameters β, this section solves the following subproblem for $P \in \{P_1, P_2, P_3\}$ in this section,

$$\text{minimize } \sum_{i=1}^{3} \sum_{X_i\in P_i} (-\lambda\lambda(X_t, \Theta_i) + \beta I\{X_{t-1} \notin P_i\}).$$

The objective of this subproblem is to assign the T subsequences to these 3 clusters based on the tradeoff of maximizing the log likelihood of the data and minimizing the volatility of cluster assignment. We can get the initial results of points assignment P_i (i=1, 2, 3) from 3^T potential assignments of points to 3 clusters through dynamic programming. The pseudo code of cluster assignment is described in Table 1.

Table 1

Cluster Assignment

Algorithm 1 Cluster Assignment

Input K←3, β←600, $-\lambda\lambda(j,i)$ that check whether point j of each influencing factors is assigned to cluster i

Output clustering assignments, initial Θ_i (i=1, 2, 3)

```
initialize PrevCost = list of 3 zeros
      CurrCost = list of 3 zeros
      PrevPath = list of 3 empty lists
      CurrPath = list of 3 empty lists
for j ← 1 to T do
  for i ← 1 to 3 do
     MinIndex = index of minimum value of PrevCost
     if PrevCost[MinIndex] +β> PrevCost[i] then
       CurrCost[i] = PrevCost[i] - λλ(j, i)
       CurrPath[i] = PrevPath[i].append[i]
    else
       CurrCost[i] = PrevCost[minIndex] +β- λλ(j, i)
       CurrPath[i] = PrevPath[minIndex].append[i]
  PrevCost = CurrCost
  PrevPath = CurrPath
FinalMinIndex = index of minimum value of CurrCost
FinalPath = CurrPath[FinalMinIndex]
return FinalPath
```

3.2.2 Toeplitz Graphical Lasso

Given the initial cluster assignments $P \in \{P_1, P_2, P_3\}$ and inverse covariances $\Theta \in \{\Theta_1, \Theta_2, \Theta_3\}$, this section solves the TICC's optimization problem and updates the parameters Θ_i and P_i (i=1, 2, 3). Based on a convergence criterion (ε=2e-5), We are able to get the global optimum of cluster assignments P_i (i=1, 2, 3) and MRF through ADMM. MRF calculated in this subsection is conducive to analyze the relative importance of each influencing factor and interpretate the cluster results in the next section. The steps of solving the Toeplitz Graphical Lasso is outlined as follows:

The pseudo code of Toeplitz Graphical Lasso is described in Table 2.

Table 2

Toeplitz Graphical Lasso

Algorithm 2 Toeplitz Graphical Lasso

Input ε←2e-5, k←100, initial Θ_i, P_i (i←1, 2, 3)

Output Θ_i, P_i (i=1, 2, 3) , MRF

repeat

$\Theta^{k+1} \leftarrow \operatorname{argmin}_{\Theta} L_p(\Theta, Z^k, U^k)$

$Z^{k+1} \leftarrow \operatorname{argmin}_{Z \in \Gamma} L_p(\Theta^{k+1}, Z, U^k)$

$U^{k+1} \leftarrow U^k + (\Theta^{k+1} - Z^{k+1})$

until $|U^{k+1} + U^k| < 2e-5$

end

Here, k is the iteration number (number of max iteration = 100), $\rho > 0$ is the penalty parameter in ADMM, $U \in R^{n \times n}$ is the scaled dual variable (Boyd et al, 2010), $Z \in \Gamma$. $L_p(\Theta, Z, U)$ is the augmented Lagrangian which can be expressed as

$$L_p(\Theta, Z, U) = -\log \det(\Theta) + \mathrm{Tr}(S\Theta) + \|\lambda \circ Z\|_1 + \frac{\rho}{2}\|\Theta - Z + U\|_F^2.$$

4. Data

The data used in this study are obtained from the Investing database maintained by the Fusion Media Limited. The data set covers time period from 07/17/2007 up to 01/09/2018.

The Hangseng Stock Index is selected as the dependent variable for the experiment. In light of the previous literature, many influencing factors in the economic environment may be used as input state variables in the construction of pattern discovery models of stock market index. Based on previous literature and the special economic situation in Hong Kong, five influencing factors are selected. Table 3 describes an array of economic variables used in this paper.

Table 3

List of Influencing Input Variables and Forecasted Output Variable

Input variables
DIA —— Dow Jones Industrial Average Index
SPX —— Standard Poor's 500 Index
IXIC —— Nasdaq Composite Index
SIR—— One-month interest rate of American Treasury which is a proxy of short-term interest rate
Dollar Index
Output variable
HSI—— Hangseng Stock Index

Many papers show that the US stock indices are leading indicators of stock indices in other stock markets, and there is bi-directional information flow between the US stock markets and other stock markets (Rahahleh, 2017; Tang *et al*., 2019). Therefore, the US stock indices may be thought of as indicators on the future level of Hangseng Stock Index. We incorporate three US stock indices in the input variables in order to discover repeated patterns of Hangseng Stock Index more efficiently. These three US stock indices include Dow Jones Industrial Average Index, Standard Poor's 500 Index and Nasdaq Composite Index.

Short-term interest rate maybe also helpful in discovering repeated patterns of stock index. When short-term interest rate is perceived as the opportunity cost of investing in equity markets, it is common to assume that the relationship between interest rate and share prices is negative. Many research has confirmed this dynamic interaction between stock returns and bond yields (Scholz *et al*., 2016). Therefore, one-month interest rate of American Treasury which is a proxy of short-term interest rate is also included in the dataset on stock index pattern discovery.

Moreover, Dollar Index may also help to discover repeated patterns of Hangseng Stock Index. Since 1983, Hong Kong implemented a currency board mechanism and pegged its dollar to the U.S. dollar. Through this fixed exchange rate system, the monetary policy in the US is of great influence on the economic situation in Hong Kong (Wong, 2019). Therefore,

dollar index as a proxy of American exchange rate may be thought of as an important indicator on the future level of Hangseng Stock Index which then indirectly, result in some power to discover repeated patterns of Hangseng Stock Index.

5. Experimental Results

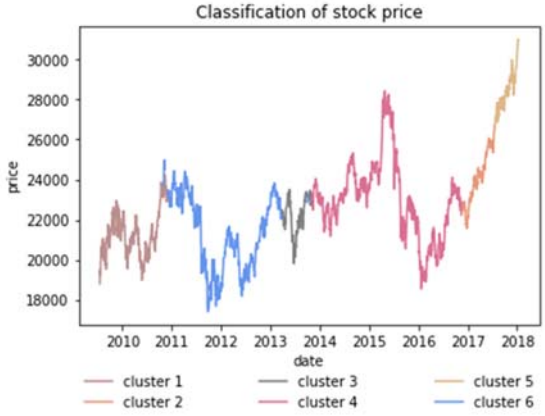
5.1 TICC Results

To discover repeated patterns of Hangseng Stock Index, we apply TICC method to pattern discovery of stock index. We start by randomly initialize the cluster parameters and cluster assignments. From there, we combine cluster assignments and the updates of cluster parameters into one EM algorithm. We repeat the E and M-steps until convergence. The results and discussion of the experiment is described in the following sections.

In the pattern discovery of Hangseng Stock Index based on TICC, a cluster number of 6 is found to produce the best possible results. W, λ, β and ε are arbitrarily chosen to be 3, $11e^{-3}$, 600 and $2e^{-5}$, respectively. The program is constructed using python 2.7 language. Figure 1 shows the result of the pattern discovery of Hangseng Stock Index.

Figure 1

Pattern Discovery of Hangseng Stock Index



From Fig. 1, one may see that six kinds of patterns are discovered in the Hangseng Stock Index, while clusters 1 and 3 with brown marking and yellow marking repeats twice respectively, and cluster 6 with blue marking repeats three times.

Specifically, the brown marking is cluster 1 which is similar to the technical analysis of “rising flag”. In the ”rising flag” pattern, investor are often rational and expect stock prices to increase. Stock price experiences a short-time soar and a slight downward trend which is followed by frequent fluctuations and a final increase. The blue marking is cluster 6 which is similar to the technical analysis of “W” shape. In the “W” shape, stock market rebounds before the end of downward trend and then fell again slightly. Finally, stock price will stop at the previous low point and begin to rise. When the stock price fell back to the last low, investors begin to take long positions and make profit. When more and more investors take long positions, the demand for this stock will drive the stock price to rise, and the stock price will break through the last high point. Therefore, in this pattern, investors could analyze

financial market carefully and buy stocks in a timely manner to make capital gains. The grey marking is cluster 3 which is similar to the technical analysis of a "V" shape. In the shape of "V", the great power of seller in the market make stock price continue to fall with stability. When the selling power disappears, the buyer's power completely controls the entire market and experiences a dramatic rebound of stock price at the same rate of decline. In the rising or falling stage of the V-shaped trend, horizontal area exists. This is because during the formation of this trend, some investors have no confidence in the form. When this power is digested, the stock price continues to complete the whole form. In the area where the V-shaped trend is extended, we can buy at the low point of this area and wait for the completion of the entire form. The red marking is cluster 4 which is similar to the technical analysis of "head and shoulders" shape. In the "head and shoulders" shape, stock price presents a "mountain" shape which is a typical signal of dramatic falls. A peak of stock price is followed by a power of resistance which leads to a failure of continuing to a new high. When the neckline of the "head and shoulders" shape is broken, a real sell signal occurs. Although compared with the highest point, the stock price has fallen back to a considerable extent, the decline trend is only at the beginning and the unsold investors could continue to sell. The orange marking is cluster 2 which is similar to a shape of "check". In the shape of "check", the stock price will experience a slight falling at beginning with the shake of investor' confidence. When buyer power increases, the stock price will increase dramatically. Investors could continue to take long positions when short decrease occurs. The yellow marking is cluster 5 which is similar to a shape of "spine". In the shape of "spine", stock price will go straight up to the highest point and then plummet. Speculators need to be aware of stock market and sell stocks in time to avoid suffering losses. A summary of Hangseng Stock Index pattern description is shown in Table 4.

Table 4

Summary of Hangseng Stock Index pattern description

Pattern number	Marking color	Shape assumed	Investor suggestion
1	brown	"rising flag"	take long position
2	orange	"check"	take long position
3	grey	"V"	choose buy timing carefully
4	red	"head and shoulders"	stop loss in time with long position or take short position
5	yellow	"spine"	make gains and stop losses in a timely manner
6	blue	"W"	choose buy timing carefully

By applying TICC algorithm to pattern discovery of Hangseng Stock Index, we find different types of repeated patterns which are similar to the graph research of technical analysis. However, unlike technical analysis, the repeated patterns we discover through TICC algorithm are more objective and comprehensive. Moreover, each repeated pattern we discovered has a flexible and long-time span, which maybe more helpful for the further research on pattern prediction and portfolio construction.

5.2 MRF and Betweenness Centrality Score

In this paper, the TICC algorithm implements pattern discovery using a correlation network or MRF defined on the short window of size w (*i.e*. 3). The MRF reflects the (time-invariant) partial correlation structure of any window within a subsequence belonging to the cluster.

In this section, in order to analyze the repeated patterns of Hangseng Stock Index further, we use network analytics to analyze the relationship between repeated patterns of Hangseng Stock Index and five influencing factors. Through this analysis, we could determine how "important" each influencing factor is in the cluster's network. Figure 2 and Table 5 show the MRF of each cluster and the betweenness centrality score of each influencing factor (Brandes, 2001), respectively.

Figure 2

Markov Random Field (MRF) of Six Clusters

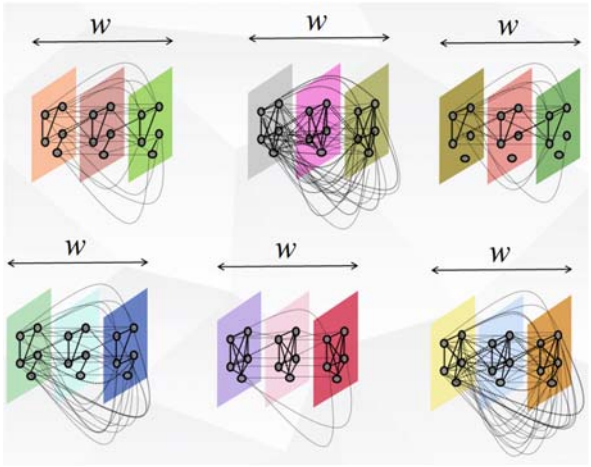


Table 5

Betweenness centrality for each influencing factor

Cluster number	Interpretation	DIA	SPX	IXIC	Dollar Index	SIR
1	Rising Flag	3.000	3.000	3.000	2.000	2.000
2	Check	5.050	2.050	15.716	2.467	15.716
3	V	4.000	4.000	4.000	1.000	0.000
4	Head and Shoulders	2.917	2.917	15.810	30.083	1.833
5	Spine	41.663	2.611	13.038	14.196	22.492
6	W	3.800	14.666	3.800	14.467	2.467

From Figure 2, we find that cluster 2 and 6 reflect a closer relationship among different influencing factors, while cluster 3 and 5 show the opposite results. In addition, we also find that different influencing factors have different betweenness scores in each cluster as shown in Table 5.

Specifically, each influencing factor has a non-zero score in cluster 1 which has an interpretation of "rising flag" from figure 1. In cluster 1, Dow Jones Industrial Average Index, Standard Poor's 500 Index, Nasdaq Composite Index are more important than the other influencing factors. In cluster 2 with an interpretation of "check", Nasdaq Composite Index

and one-month interest rate of American Treasury has the largest betweenness score. In cluster 3 with an interpretation of "V" shape, one-month interest rate of American Treasury has no influence on the pattern of Hangseng Stock Index while Dow Jones Industrial Average Index, Standard Poor's 500 Index, Nasdaq Composite Index have the same largest betweenness score. Cluster 4 is the cluster with the least importance on one-month interest rate of American Treasury and the largest Dollar Index score, which is interpreted as "Head and Shoulders" in Figure 1. In cluster 5 with an interpretation of "spine", Nasdaq Composite Index has the largest influential power on the pattern of Hangseng Stock Index. In cluster 6 with an interpretation of "W" shape, one-month interest rate of American Treasury has the smallest betweenness score while Standard Poor's 500 Index has the largest influence on the pattern of Hangseng Stock Index.

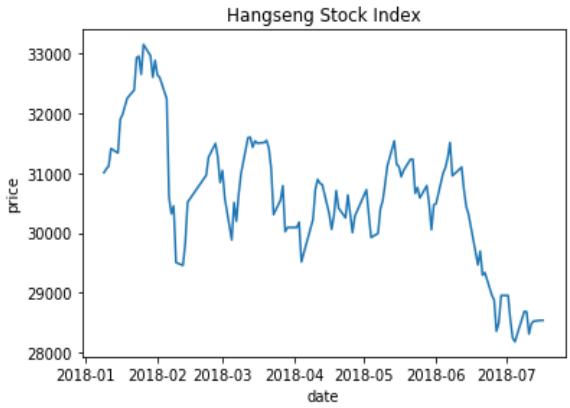
Through MRFs and betweenness scores, we could link the interpretation of each cluster in Figure 1 with the relative importance of each influencing factor, which may be helpful for future research on pattern prediction.

5.3 Robustness Analysis

Following Hallac *et al*. (2017), we discover six repeated patterns of Hangseng Stock Index based on TICC method with five variables: Dow Jones Industrial Average Index, Standard Poor's 500 Index, Nasdaq Composite Index, One-month interest rate of American Treasury, and Dollar Index. In order to confirm the feasibility of pattern discovery through TICC method, we further analyze the repeated pattern of Hangseng Stock Index from 01/09/2018 to 7/17/2018 which is shown in Figure 3.

Figure 3

Price Movement of Hangseng Stock Index



As shown in Figure 3, the repeated patterns we discover through TICC method are consistent with the price movement in realistic stock market. Specifically, the latest repeated pattern before 01/09/2018 is assumed as "spine" in which stock price will experience a final plummet. This assumption is consistent with the reality that stock price experiences a nosedive from 01/09/2018 to 7/17/2018.

6. Conclusions

Pattern discovery provides a fundamental aid for the comprehension of stock index and, more specifically, is crucial in financial planning. In the literature on pattern discovery analysis and multivariate time series analysis of stock index, scholars mainly focus on the relationship between single stock's current prices and past prices without considering influencing factors, and moreover, generally rely on distance-based metrics. In this paper, we apply a rather different approach: Toeplitz Inverse Covariance-based Clustering (TICC) to discover repeated patterns with flexible time periods in stock index. TICC is a new type of graph-based clustering method that is able to discover the relationship between repeated patterns of stock index and relative importance of each influencing factor.

This study applies the TICC method to discover repeated patterns of Hangseng Stock Index's multivariate time series and, moreover, investigate the relative importance of five influencing factors in each cluster through MRF and betweenness centrality score. The main results in the empirical research can be summarized as follows. First, six kinds of patterns are discovered in Hangseng Stock Index through Toeplitz Inverse Covariance-based Clustering (TICC) method. In pattern 1 and pattern 2 which are similar to the technical analysis of "rising flag" and "check" respectively, investors could take long positions in Hangseng Stock Index. In pattern 3 and pattern 6 which are in the shape of "V" and "W" respectively, investors could take long positions with caution in stock index. Moreover, in pattern 4 and 5 which are similar to the technical analysis of " head and shoulders" and "spine" respectively, we suggest investors and speculators to stock losses in a timely manner or take short positions in stock index.Second, the analysis of five influencing factors through MRF and betweenness centrality scores shows that the economic environment in the United States has huge influence on the stock market in Hong Kong, and that different influencing factors have different betweenness scores in each cluster.

With consideration of the interdependency among different influencing factors and the time-invariant structure in each cluster over short subsequences, we use the TICC method to discover six clusters of Hangseng Stock Index with flexible and long-time span. Moreover, we analyze the relationship between each pattern and the influencing power of each factor through MRF and betweenness centrality scores. This empirical research fills the gap between financial analysis and pattern discovery through machine learning methods. More specifically, the flexible repeated patterns we discover through the TICC method could reduce the risks of overfitting, which maybe more suitable to other stock indexes analysis. In addition, the relationship between long-period repeated patterns and different influencing factors that are discovered in this paper can provide aid for future research on pattern prediction and portfolio construction.

However, there still are some limitations in this paper. First, the relationship between different patterns and influencing factors is analyzed roughly. Second, the robustness test is very simple to verify the feasibility of pattern discovery through TICC method. Therefore, two possible extensions of stock index pattern analysis exist in future research. A possible extension of stock index pattern analysis is to investigate the relationship between different patterns and influencing factors' betweenness centrality score deeply, which maybe helpful for future research on pattern prediction. The other is to analyze whether trading strategies guided by forecasts of repeated patterns is more effective and lead to higher profits, which can provide more robustness for pattern discovery through the TICC method in this paper.

References

Al Rahahleh, N. Bhatti, M.I. and Adeinat, I., 2017. Tail dependence and information flow: Evidence from international equity markets. *Physica A: Statistical Mechanics and its Applications*, 474, pp. 319-329.

Akita, R. Yoshihara, A. Matsubara, T. and Uehara, K., 2016. Deep learning for stock prediction using numerical and textual information. *In 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, pp.1-6.

Arévalo, R. García, J. Guijarro, F. and Peris, A., 2017. A dynamic trading rule based on filtered flag pattern recognition for stock market price forecasting. *Expert Systems with Applications*, 81, pp.177-192.

Avramov, D. and Chordia, T., 2006. Asset pricing models and financial market anomalies. *The Review of Financial Studies*, 19(3), pp.1001-1040.

Bannigidadmath, D. and Narayan, P.K., 2016. Stock return predictability and determinants of predictability and profits. *Emerging Markets Review*, 26, pp.153-173.

Baumeister, T.R. Kolind, S.H. MacKay, A.L. and McKeown, M.J., 2020. Inherent spatial structure in myelin water fraction maps. *Magnetic resonance imaging*, 67, pp. 33-42.

Bouri, E. Gupta, R. Hosseini, S. and Lau, C.K.M., 2018. Does global fear predict fear in BRICS stock markets? Evidence from a Bayesian Graphical Structural VAR model. *Emerging Markets Review*, 34, pp.124-142.

Boyd, S. Parikh, N. Chu, E. Peleato, B. and Eckstein, J., 2010. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends in Machine Learning*, 3(1), pp.1-122.

Brandes, U., 2001. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25(2), pp.163-177.

Cen, X. and Qin, J., 2016. Application of Improved k-means Clustering on Trend of Stock Price Fluctuation. *Technology and Industry*, 16(1), pp.144-148.

Chen, A.-S. Leung, M.T. and Daouk, H., 2003. Application of neural networks to an emerging financial market: forecasting and trading the Taiwan Stock Index. *Computers and Operations Research*, 30(6), pp.901-923.

Chen, T.L. and Chen, F.Y., 2016. An intelligent pattern recognition model for supporting investment decisions in stock market. *Information Sciences*, 346, pp.261-274.

Chou, J. S. and Ngo, N.T., 2016. Time series analytics using sliding window metaheuristic optimization-based machine learning system for identifying building energy consumption patterns. *Applied energy*, 177, pp.751-770.

E. Parzen, 1962. On estimation of probability density function and mode. *Annals of Math. Statistics*, 33, pp.1065-1076.

Fu, T.C. Chung, F. L. Ng, V. and Luk, R., 2001. Pattern discovery from stock time series using self-organizing maps. *In Workshop Notes of KDD2001 Workshop on Temporal Data Mining*, pp. 26-29.

Friedman, J. Hastie, T. and Tibshirani, R., 2008. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3), pp. 432-441.

Gionis, A. and Mannila, H., 2003. Finding recurrent sources in sequences. *In Proceedings of the seventh annual international conference on Research in computational molecular biology*, pp.123-130.

Guo, N. Xiao, R. Gao, S. and Tang, H., 2017. A neurally inspired pattern recognition approach with latency-phase encoding and precise-spike-driven rule in spiking neural network. *IEEE International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics, Automation and Mechatronics (RAM)*, pp.19-21.

Hallac, D. Vare, S. Boyd, S. and Leskovec, J., 2017. Toeplitz Inverse Covariance-Based Clustering of Multivariate Time Series Data. *The 23rd ACM SIGKDD International Conference*, pp. 215-223.

Jhaveri, R.H. Patel, N.M. Zhong, Y. and Sangaiah, A.K., 2018. Sensitivity analysis of an attack-pattern discovery based trusted routing scheme for mobile ad-hoc networks in industrial IoT. *IEEE Access*, 6, pp. 20085-20103.

Kahneman, D. and Tversky, A., 2013. Prospect theory: An analysis of decision under risk. *In Handbook of the fundamentals of financial decision making: Part I* , pp. 99-127.

Kazem, A. Sharifi, E. Hussain, F.K., Saberi, M. and Hussain, O.K., 2013. Support vector regression with chaos-based firefly algorithm for stock market price forecasting. *Applied Soft Computing*, 13(2), pp. 947-958.

Koller, D. and Friedman, N., 2009. Probabilistic graphical models: principles and techniques. *NJ: MIT Press.*

Li, F. Sheng, H. and Zhang, D., 2002. Event Pattern Discovery from the Stock Market Bulletin. *International Conference on Discovery Science*, 2534, pp. 310-315.

Lo, A.W., 2004. The adaptive markets hypothesis. *The Journal of Portfolio Management*, 30(5), pp.15-29.

Luo, L. and Chen, X., 2013. Integrating piecewise linear representation and weighted support vector machine for stock trading signal prediction. *Applied Soft Computing*, 13(2), pp. 806-816.

Malinowski, S. Guyet, T. Quiniou, R. and Tavenard, R., 2013. 1d-SAX: A Novel Symbolic Representation for Time Series. *Advances in Intelligent Data Analysis*, 8207 (12), pp. 273-284.

Malkiel, B.G. and Fama, E.F., 1970. Efficient capital markets: A review of theory and empirical work. *The journal of Finance*, 25(2), pp. 383-417.

McLean, R.D. and Pontiff, J., 2016. Does academic research destroy stock return predictability?. *The Journal of Finance*, 71(1), pp. 5-32.

Min, F. Zhang, Z.H. Zhai, W.J. and Shen, R.P., 2020. Frequent pattern discovery with tri-partition alphabets. *Information Sciences*, 507, pp. 715-732.

Moghaddam, A.H. Moghaddam, M.H. and Esfandyari, M., 2016. Stock market index prediction using artificial neural network. *Journal of Economics, Finance and Administrative Science*, 21(41), pp. 89-93.

Nisar, T.M. and Yeung, M., 2018. Twitter as a tool for forecasting stock market movements: A short-window event study. *The journal of finance and data science*, 4(2), pp.101-119.

Scholz, M. Sperlich, S. and Nielsen, J.P., 2016. Nonparametric long term prediction of stock returns with generated bond yields. *Insurance: Mathematics and Economics*, 69, pp. 82-96.

Schumaker, R.P. Labedz Jr, C.S. Jarmoszko, A.T. and Brown, L.L., 2017. Prediction from regional angst–A study of NFL sentiment in Twitter using technical stock market charting. *Decision Support Systems*, 98, pp. 80-88.

Sezer, O.B. Ozbayoglu, M. and Dogdu, E., 2017. A deep neural-network based stock trading system based on evolutionary optimized technical analysis parameters. *Procedia computer science*, 114, pp. 473-480.

Shen, S. and Shen, Y., 2016. ARIMA model in the application of Shanghai and Shenzhen stock index. *Applied Mathematics*, 7(3), pp. 171-176.

Steffen L. Lauritzen, 1996. Graphical Models. *Statistics in Medicine*, 18(21), pp. 2983-2984.

Tang, Y. Xiong, J.J. Luo, Y. and Zhang, Y.C., 2019. How Do the Global Stock Markets Influence One Another? Evidence from Finance Big Data and Granger Causality Directed Network. *International Journal of Electronic Commerce*, 23(1), pp. 85-109.

Wang, J.L. and Chan, S.H., 2007. Stock market trading rule discovery using pattern recognition and technical analysis. *Expert Systems with Applications*, 33(2), pp. 304-315.

Wieldraaijer, T. Bruin, P. Duineveld, L.A. Tanis, P.J. Smits, A.B. van Weert, H.C. and Wind, J., 2018. Clinical pattern of recurrent disease during the follow-up of rectal carcinoma. *Digestive surgery*, 35(1), pp. 35-41.

Wong, E.M., 2019. A Comparison of the Economic Volatility Spillover Effect of Hong Kong with China and USA. *Asian Economic and Financial Review*, 9(7), pp. 824.