International Journal of Agricultural and Environmental Information Systems
Volume 10 • Issue 4 • October-December 2019

Soybean Price Pattern Discovery Via Toeplitz Inverse Covariance-Based Clustering

Hua Ling Deng, Northeast Agricultural University, Harbin, China
Yǔ Qiàn Sūn, Northeast Agricultural University, Harbin, China

ABSTRACT

The high volatility of world soybean prices has caused uncertainty and vulnerability particularly in the developing countries. The clustering of time series is a serviceable tool for discovering soybean price patterns in temporal data. However, traditional clustering method cannot represent the continuity of price data very well, nor keep a watchful eye on the correlation between factors. In this work, the authors use the Toeplitz Inverse Covariance-Based Clustering of Multivariate Time Series Data (TICC) to soybean price pattern discovery. This is a new method for multivariate time series clustering, which can simultaneously segment and cluster the time series data. Each pattern in the TICC method is defined by a Markov random field (MRF), characterizing the interdependencies between different factors of that pattern. Based on this representation, the characteristics of each pattern and the importance of each factor can be portrayed. The work provides a new way of thinking about market price prediction for agricultural products.

KEYWORDS

High Volatility, Multivariate Analysis, Pattern Discovery, Time Series Clustering

INTRODUCTION

The soybean is one of the requisite grain crops in the world, and has been cultivated for more than 5,000 years. Due to its high nutritive value, amounts of full-fat soybeans are being used in the feed and food industry. In addition, since the soybean, as the vital economic crop, plays an important role in economics and trade around the world, the

DOI: 10.4018/IJAEIS.2019100101

soybean price index has become an important indicator of China's economic activity. The instability of soybean prices will bring huge risks to farmers, governments, consumers, and other commercial entity involved in soybean market. Therefore, accurate analysis of soybean market price needs to be taken seriously.

There is a vast literature on analysis of agricultural product price. On the one hand, several studies examine the relationship between agricultural product price and other factors. For example, Harri et al. (2009) investigated the cointegration relationship between exchange rates, oil prices, and agricultural crop prices by k-*th* order Vector Autoregression (VAR) model. Nazlioglu et al. (2012) studied the dynamic relationship between world oil prices and twenty-four world agricultural commodity prices by panel cointegration and Granger causality methods. Ekananda et al. (2018) observed that the world soybean price and exchange rate may affect the domestic soybean prices positively and significantly in the short term by Bound Testing Cointegration method with Autoregressive Distributed Lag (ARDL) approach. On the other hand, some researchers prefer to fit time series by selecting an appropriate model to predict agricultural product price. For example, Assis et al. (2010) and Maizah et al. (2014) predicted Cocoa Bean price sequences and the prices of Malaysian crude palm oil by Generalized AutoRegressive Conditional Heteroskedasticity (GARCH) model respectively. Octavio et al. (2009) achieved a better predicted result of U.S. soybeans and Brazilian coffee prices by Threshold Autoregressive (TAR) model. The prediction methods above can provide some valuable information for decision makers. However, these methods either cannot overcome sensitivity to noise, or only forecast in a short-term. To solve noisy sensitivity and short-term forecasting, agricultural product price sequence can be divided into several subsequences, in which each subsequence belongs to some defined trend or "pattern" reoccurring in the future. Once these patterns come under observation, seemingly unordered price data can be interpreted as a few defined patterns. The process of finding pattern is referred to "pattern discovery". Pattern discovery try to forecast the trend of agricultural product price rather than the price in short-term. Due to the noise of data has less influence on the trend prediction than the short-terming forecasting, pattern discovery can overcome sensitivity to noise and forecast the trend of agricultural product price (Gionis et al., 2003).

Pattern discovery from time series is of fundamental importance. Particularly with the development of sensors, time series has become an important class of temporal data objects and they can be easily obtained from many applications, e.g., daily temperatures, levels of pollution, human heartbeats, and prices of agricultural products. Different from the traditional discrete database, time series data are characterized by their continuity. Therefore, when they can be focused as fragments rather than as individual data points, interesting patterns can be discovered.

For the problem of pattern discovery in time series, one of the most popular techniques being employed is clustering since clustering methods can seek out the similarity and distinction between data, namely, different patterns in the data. Most previous studies in time series clustering are restricted to univariate time series. For example, Golay et al. (1998) applied the fuzzy c-means algorithm to univariate time

series in order to study the human brain activity. Tak-chung Fu et al. (2001) proposed the self-organizing map (SOM) based pattern discovery scheme that cooperates with a new pattern matching scheme for discover the pattern of stock price. They tested SOM algorithms on a few Hong Kong Stock Exchange listed stocks' time series. In practical problems, however, multivariate time series data are actually complex and related: when a car is moving, for example, the data of sensors such as Gas Pedal Position, Vehicle Velocity and Brake Pedal Position may codetermine whether the current car are turning or speeding up. These characteristics makes various algorithms have been proposed to cluster multivariate time series data of different types (Liao, 2005).

The algorithms that addressed multivariate time series include Košmelj et al. (1990), Kakizawa et al. (1998), Ramoni et al. (2000), etc. Most of these studies assume that there is no cross-correlation between factors and simplify the overall joint distribution by assuming conditional independence between factors. Moreover, these methods have no idea to explain the cluster results. In the conference on Knowledge Discovery and Data Mining, Hallac et al. (2017) proposed a new method of model-based clustering, the Toeplitz Inverse Covariance-based Clustering (TICC), which is the main method we are going to use. Hallac defined each cluster by a correlation network, or Markov random field (MRF), describing the interdependencies between different observations. In this manner, TICC can output the key factors affecting each category by MRF which is also the unique in this approach. Procacci et al. (2018) identified the state of financial markets based on the TICC model. Their experiment is equally proof that the method is efficient and reliable in identifying and predicting accurate and interpretable structures in multivariate, non-stationary financial datasets. To the best of our knowledge, our research is the first to apply TICC method to the field of agricultural price analysis.

In this paper, a clustering methodology based on Toeplitz matrix is proposed for discover the agricultural products market price pattern. As an application, this study evaluated the price pattern of soybean market over the Heilongjiang province. The remainder of this paper is organized as follows: In Sect. 2, we mainly introduce the mathematical model which used in this article. In section 3, we show how variables and data are selected based on our experimental. In Sect. 4, we offer a detailed analysis and empirical evaluation of our ideas. Finally, in Sect. 5, we provide conclusions and directions for future work.

MATHEMATICAL MODEL ON SOYBEAN PRICE PATTERN DISCOVERY

Just like static data clustering, the choice of multivariate time series clustering algorithm depends both on the characteristics of data available and the practical problem. In the most case, the market price of soybean will be also affected by other factors (such as the purchase price of soybean, the futures price of soybean oil, etc.), which forms a set of multivariate time series data. These data are characterized by continuity, correlation, and time-invariant. To achieve this clustering, it is necessary

to simultaneously segment and cluster multivariate data. This seems more difficult than traditional time series segmentation, since multiple segments can belong to the same cluster.

Compared with the traditional clustering, TICC is a new type of model-based multivariate time series clustering method, which can find the accurate and interpretable structure in the data. TICC describes each cluster with different MRFs and solved the problem of simultaneous segmentation and clustering through alternating minimization, using a variation of the expectation maximization (EM) algorithm.

For simplicity of notation, we consider a time series of T sequential observations, $x = [x_1, x_2, x_3, \cdots x_T]$, where $x_i \in R^n$ is the i-th n-dimensional time series data. However, instead of looking at x_t separately, we intercept a short subsequence of size w though TICC algorithm is relatively robust to the selection of this window size parameter, $x_i = [x_{i-w+1}, \cdots x_{i-1}, x_i]$, where $x_i \in R^{nw}$. We take this new sequence as $X = [X_1, X_2, X_3, \cdots X_T]$. The operation of the TICC can cluster these T observations into K clusters or states, where the set of observations of cluster j is denoted as $P_j, j = 1,2,\cdots k$. In addition, each cluster defined by a Gaussian inverse covariance $\Theta \in R^{nw \times nw}$ or MRF, describing the interdependencies between different variates in a typical time sequence of that cluster. Note that Θ_i is composed of $w \times w$ submatrix, each of which is $n \times n$. Here, the submatrix of the PQ position describes the covariance inverse matrix between time P and time Q. In other words, Θ_i is a partitioned Toeplitz matrix that can be expressed in the following form:

$$\Theta_i = \begin{bmatrix} A_0 & A_1^T & A_2^T \\ A_1 & A_0 & A_1^T \\ A_2 & A_1 & A_0 \end{bmatrix}$$

TICC cluster by solving two key problems in turn: Assign points to clusters, where we will use a dynamic programming (DP) algorithm, and Up-date Cluster Parameters, where we solve the Toeplitz graphical lasso problem using an algorithm based on the alternating direction method of multipliers (ADMM).

Assign Points to Clusters

Given Θ_j, the cost of assigning X_t into cluster j is equivalent to the negative logarithmic likelihood:

$$E(i \in P_j) = -ll(i,j) = -\log[N(X_i); 0, \Theta_j^{-1}]$$

$$= -\log\left[\det\left(\Theta_j\right)^{1/2} \cdot \exp\left[-\frac{1}{2} X_i^T \Theta_j X_i\right]\right] = -\log\det\left(\Theta_j\right) + X_i^T \Theta_j X_i$$

In addition, considering the continuity of the observations, smoothness penalty parameter β is imposed when the adjacent data is not belong to the same cluster:

$$E(i,i+1)=\begin{cases}0 & i,i+1 \quad \text{same cluster}\\ \beta & i,i+1 \quad \text{not the same}\end{cases}$$

A typical Assembly-Lines-Scheduling problem is comprised by these two costs, which can be solved efficiently by dynamic programming algorithm. Dynamic programming (usually referred to as DP) is a powerful technique for solving complex problems by decomposing them into a set of simpler subproblems, solving each of those subproblems just once, and using memory-based data structures (arrays, maps, and so on) to store their solutions. (Bellman, 1957). The core of the algorithm lies on only considering the cost of the $i-1$th observation value when assigning the ith observations.

Algorithm 1. Assign points to clusters

```
given β > 0, −ℓℓ(i, j) parameters
initialize PrevCost = list of K zeros.
           CurrCost = list of K zeros.
           PrevPath = list of K empty lists.
           CurrPath = list of K empty lists.
for i = 1, . . ., T do
      for j = 1, . . ., K do
            MinIndex = index of minimum value of PrevCost.
            if PrevCost[MinIndex] + β > PrevCost[j] then
                CurrCost[j] = PrevCost[j] −ℓℓ(i, j).
                CurrPath[j] = PrevPath[j].append[j].
      else
                CurrCost[j] = PrevCost[minIndex] + β − ℓℓ(i, j).
                CurrPath[j] = PrevPath[minIndex].append[j].
      PrevCost = CurrCost.
      PrevPath = CurrPath.
FinalMinIndex = index of minimum value of CurrCost.
FinalPath = CurrPath[FinalMinIndex].
return FinalPath.
```

Update Cluster Parameters

Once each point has been assigned, we can update the inverse covariances of each cluster by minimizing its negative logarithm likelihood summation, which can be written as follows:

$$E(\Theta) = \sum_{i \in P} -ll\left(X_i, \Theta\right) = \sum_{i \in P} -\log\det(\Theta) + X_i^T \Theta X_i = E_1 + E_2$$

Here, E_1 is independent of j and can be rewritten as its equivalent problem:

$$E_1 = -|P_i| \cdot \log\det(\Theta)$$

where $||$ represents the number of elements in a cluster. Similarly, E_2 can be rewritten as the trace:

$$E_2 = -tr\left(\sum_{i \in P} X_i^T X_i\right) = -\left|P_i\right| \cdot tr\left(S \cdot \Theta\right)$$

where S is the covariance matrix calculated by P_j. Add another one regularization parameters: λ, which determines the sparsity level in the MRFs characterizing each cluster:

$$E_3 = || \lambda \circ \Theta ||_1$$

where λ is the weight matrix, $\circ$ is the matrix multiplication, and Θ is a block Toeplitz matrix (Gray, 2011).

It is necessary to develop a fast method for solving it efficiently. By transforming the problem into the following form, we can use the alternating direction method of multipliers (ADMM) to achieve this idea, a distributed convex optimization approach that has been shown to perform well at large-scale optimization tasks (Boyd et al., 2011; Parikh at al., 2014):

$$\min imize - \log\det\Theta + tr(S \cdot \Theta) + \left\|\lambda \circ Z\right\|_1$$

$$subject\ to \quad \Theta = \mathrm{Z}$$

where Z is a block Toeplitz matrix.

Algorithm 2. Update cluster parameters

```
initialize Cluster parameters Θ; cluster assignments P.
repeat
        E-step: Assign points to clusters → P.
until Stationarity.
return (Θ, P).
```

We can implement the TICC algorithm by a Python solver.[1] TICC solver breaks the T timestamps into segments where each segment belongs to one of the "k" clusters. The total number of segments is defined by the smoothness parameter "β". It does so by running an EM algorithm where TICC alternately assigns points to clusters using a DP algorithm and updates the cluster parameters by solving a Toeplitz inverse covariance estimation problem. In short, the solver takes as input a T-by-n data matrix, the window size "w", the number of clusters "k" and some necessary parameters, then returns an array of cluster assignments for each time point. Simultaneously, we can get a dictionary with keys being the cluster_id (from 0 to k-1) and the values being MRFs of each cluster. Below, we will demonstrate our experiment and analysis in two sections.

EXPERIMENTS

Heilongjiang accounts for about 60 percent of China's soybean acreage, producing 1.5 million tonnes a year, about 11% of the country. In this section, we attempt to model the common patterns of soybean price in Heilongjiang province. In Hallac's literature, TICC method has shown good performance when compared with several state-of-the-art baselines such as Gaussian Mixture Model (GMM) and Dynamic Time Warping (DTW). To provide a complete evaluation of the models, Hallac validated TICC approach by comparing TICC to several state-of-the-art baselines in a series of synthetic experiments and an automobile sensor dataset respectively. And the finding was that compared to other several well-known time series clustering approaches, TICC not only has high accuracy, but also significantly outperforms the baselines in scalability and interpretability. For these reasons, our algorithms which named Toeplitz Inverse Covariance-based Clustering (TICC) can be applied to find price pattern properly. The main experimental steps are as follows:

Step 1: Choose the factors.
Step 2: Data acquisition and processing.
Step 3: Run the TICC to cluster.
Step 4: Comprehensive analysis.

According to previous research, quite a few factors will affect the change of soybean market price together due to their different characteristics and the requirement of society (Liu et al., 2005; Massimo et al., 2013; Wang, 2016). We finally selected 4 factors as the independent variables of the experiment after tests and analysis: soybean purchase price, corn market prices, soybean and soybean-oil futures prices. We cluster the multiple time series and map the clustering results to the soybean market price, and we will find that there are truly similar patterns to repeat in some periods.

Our database is obtained from the Dalian Commodity Exchange (http://www.dce.com.cn/DCE/DCE_PAGE_KEY/index.html), Heilongjiang Agricultural Information (http://www.hljagri.gov.cn/ddw/scbj/), the Zhujiage network (http://www.zhujiage.

com.cn/special/hlj_dadoujiage.html) by writing a web crawler which is a program can automatically retrieve web content. Here, it is worth taking a moment to flag the fact that the data on the page could be partially missing. To get through this, we used Lagrange interpolation method to fill the missing data. Lagrange interpolation is a mathematical method which tries to construct an appropriate function that can pass through several known points, and then use the function to find the unknown points on the interval. The Scipy library in python has a Lagrange function that makes it easy to do this task.

In a previous work, we introduced the parameters of TICC, it can be selected by a precise method such as Bayesian information criterion (BIC) or cross-validation (Hastie et al, 2009). We pick the number of clusters using BIC, and we do the trick when K = 4. We run TICC with a window size of 3 day. Thus, in this experiment, we have 358 price data of a 4-dimensional time series. After a few minutes of running, we will get a CSV and a TXT file, containing an array of clustering results and a MRF for each cluster respectively, as indicated in the code. All the cases have been run on a PC with win7 of operating system and a 2.4-GHz-based processor.

RESULTS AND DISCUSSIONS

We can see in Figure1 that TICC divides the entire timeline into five segments where the first (2016.7.15-2016.9.12) and third (2017.3.28-2017.5.26) segment belong to the same cluster. Therefore, we can expect that the soybean market can be represented four patterns in this timeline, although we are insensible of what these four models represent at present. Furthermore, we notice that the two segments that are both in Pattern 1 are not in the same month. That means the patterns of soybean price is not affected by season to some extent.

Recall that we define each cluster by a sparse Gaussian inverse covariance matrix Θ which illustrates the conditional independency structure between the variables (Friedman et al., 2008). Here, we also give the definition matrix for each cluster shown in Tables 1-4.

Note that the inverse covariance matrix of a different cluster have a different value between any two factors. Actually, the number of ij positions in the matrix refers to the relationship between concurrent values of i and j. Here, the A_0 sub-block represents the intra-time partial correlations, so A_{0ij} refers to the relationship between concurrent values of factors i and j. Similarly, A_{1ij} shows how factor i at some time t is correlated to sensor j at time t + 1, and A_2 shows the edge structure between time t and time t + 2(Figure2). For example, the order of the four values is 18.1207, 3.2406, 11.4418, and 6.8134 in A_0, which represents the order of autocorrelation of variables is factor 1, 3,4, 2 (soybean purchase price, soybean futures prices, soybean-oil futures prices, and corn market prices, respectively). The number 0.071 and 0.625 means that the variable 4 has a relationship with the variable 2.3, and the variables 3 and the variable 4 are more closely related.

Figure 1. Multiple time series and clustering results. Each curve represents a set of data.

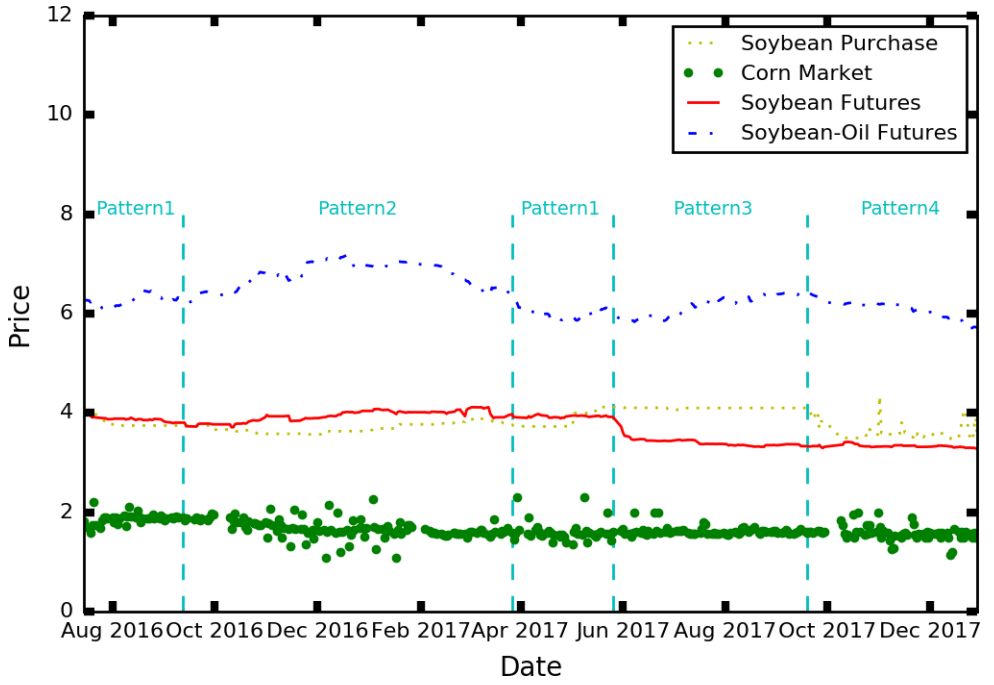


Table 1. Pattern 1

18.1207	0	0	0	0	0	0	0	0	0	0	0
0	3.2406	0	0.071	0	0	0	0.0202	0	0	0	0
0	0	11.4418	0.625	0	0	0	0.7182	0	0	0	0.8802
0	0.071	0.625	6.8134	0	0	0.5378	-2.6792	0	0	0.5476	-2.5472
0	0	0	0	18.1207	0	0	0	0	0	0	0
0	0	0	0	0	3.2406	0	0.071	0	0	0	0.0202
0	0	0	0.5378	0	0	11.4418	0.625	0	0	0	0.7182
0	0.0202	0.7182	-2.6792	0	0.071	0.625	6.8134	0	0	0.5378	-2.6792
0	0	0	0	0	0	0	0	18.1207	0	0	0
0	0	0	0	0	0	0	0	0	3.2406	0	0.071
0	0	0	0.5476	0	0	0	0.5378	0	0	11.4418	0.625
0	0	0.8802	-2.5472	0	0.0202	0.7182	-2.6792	0	0.071	0.625	6.8134

Table 2. Pattern 2

7.0698	0.0481	-0.1771	0.1068	-2.6424	0.0326	-0.1168	0.2911	-2.5251	0.0178	-0.0326	0.4273
0.0481	1.0337	0.2061	0.0485	0.0193	-0.1654	0.1721	0.1675	0.136	0.1741	0.3311	0.033
-0.1771	0.2061	7.5867	-0.0484	-0.145	0.2914	-1.9369	0	-0.2605	0.3328	-1.6327	0
0.1068	0.0485	-0.0484	6.337	0	0	-0.3121	-2.7283	0	0.1978	-0.363	-2.5075
-2.6424	0.0193	-0.145	0	7.0698	0.0481	-0.1771	0.1068	-2.6424	0.0326	-0.1168	0.2911
0.0326	-0.1654	0.2914	0	0.0481	1.0337	0.2061	0.0485	0.0193	-0.1654	0.1721	0.1675
-0.1168	0.1721	-1.9369	-0.3121	-0.1771	0.2061	7.5867	-0.0484	-0.145	0.2914	-1.9369	0
0.2911	0.1675	0	-2.7283	0.1068	0.0485	-0.0484	6.337	0	0	-0.3121	-2.7283
-2.5251	0.136	-0.2605	0	-2.6424	0.0193	-0.145	0	7.0698	0.0481	-0.1771	0.1068
0.0178	0.1741	0.3328	0.1978	0.0326	-0.1654	0.2914	0	0.0481	1.0337	0.2061	0.0485
-0.0326	0.3311	-1.6327	-0.363	-0.1168	0.1721	-1.9369	-0.3121	-0.1771	0.2061	7.5867	-0.0484
0.4273	0.033	0	-2.5075	0.2911	0.1675	0	-2.7283	0.1068	0.0485	-0.0484	6.337

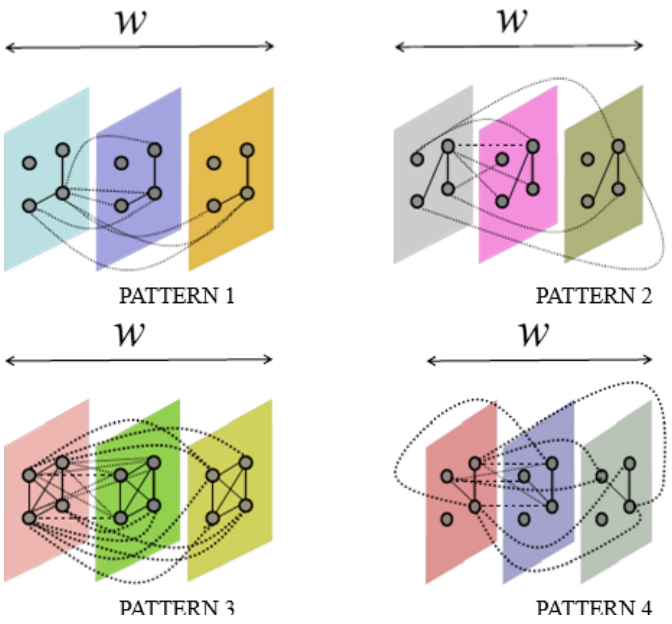
Table 3. Pattern 3

5.9649	0	0	0	-2.6506	0.0209	0	0	-2.3769	0.0455	0	0
0	1.0412	0.1651	-0.4567	0.0906	-0.1282	0.1613	-0.2051	0.1918	0	0.1218	-0.5589
0	0.1651	12.3183	0	0	0.2143	0	0	0	0.1667	0	0
0	-0.4567	0	7.2808	0	-0.494	0	-1.996	0	-0.3138	0	-1.5536
-2.6506	0.0906	0	0	5.9649	0	0	0	-2.6506	0.0209	0	0
0.0209	-0.1282	0.2143	-0.494	0	1.0412	0.1651	-0.4567	0.0906	-0.1282	0.1613	-0.2051
0	0.1613	0	0	0	0.1651	12.3183	0	0	0.2143	0	0
0	-0.2051	0	-1.996	0	-0.4567	0	7.2808	0	-0.494	0	-1.996
-2.3769	0.1918	0	0	-2.6506	0.0906	0	0	5.9649	0	0	0
0.0455	0	0.1667	-0.3138	0.0209	-0.1282	0.2143	-0.494	0	1.0412	0.1651	-0.4567
0	0.1218	0	0	0	0.1613	0	0	0	0.1651	12.3183	0
0	-0.5589	0	-1.5536	0	-0.2051	0	-1.996	0	-0.4567	0	7.2808

Table 4. Pattern 4

1.5632	0	0	-0.0399	-0.4837	0	0	0	-0.0619	-0.0201	0	0
0	1.3569	0	-0.0749	-0.0618	-0.2807	0	-0.1024	0	0	0	-0.0373
0	0	12.0629	0	0	0	0	0	0	0	0	0
-0.0399	-0.0749	0	5.7111	0	-0.053	0	-2.1033	0	-0.0696	0	-1.2704
-0.4837	-0.0618	0	0	1.5632	0	0	-0.0399	-0.4837	0	0	0
0	-0.2807	0	-0.053	0	1.3569	0	-0.0749	-0.0618	-0.2807	0	-0.1024
0	0	0	0	0	0	12.0629	0	0	0	0	0
0	-0.1024	0	-2.1033	-0.0399	-0.0749	0	5.7111	0	-0.053	0	-2.1033
-0.0619	0	0	0	-0.4837	-0.0618	0	0	1.5632	0	0	-0.0399
-0.0201	0	0	-0.0696	0	-0.2807	0	-0.053	0	1.3569	0	-0.0749
0	0	0	0	0	0	0	0	0	0	12.0629	0
0	-0.0373	0	-1.2704	0	-0.1024	0	-2.1033	-0.0399	-0.0749	0	5.7111

Figure 2. TICC method segments a time series into a sequence of states, or "clusters" (i.e., 1, 2, 3 or 4). Each cluster is characterized by a correlation network, or MRF, defined over a short window of size w. This MRF governs the (time-invariant) partial correlation structure of any window inside a segment belonging to that cluster. When $w = 3$, the visual diagram structure of MRFs is shown in the figure. The dotted line in the figure indicates that there has some relationship between the two nodes.



In order to analyze the soybean market pattern, we shall presently review some of the economic theories. The phenomenon of recurrent fluctuation between boom and depression in the economic activity is called Business Cycle (or Trade Cycle). Generally speaking, a cycle consists of four stages: Boom, Recession, Depression and Recovery. In which that Recession is a transition from Boom to Depression, the recovery is the phase of the transition from Depression to Boom. According to the theory of Business Cycle, it is inevitable that economic expansion and contraction will occur periodically. In particular, when the economy begins to falter, commodity prices fall and so does market demand. When economies reach prosperity, prices rise and production increases rapidly (Edward, 1986). Similarly, soybean price fluctuations can be explained well by the Business Cycle theory.

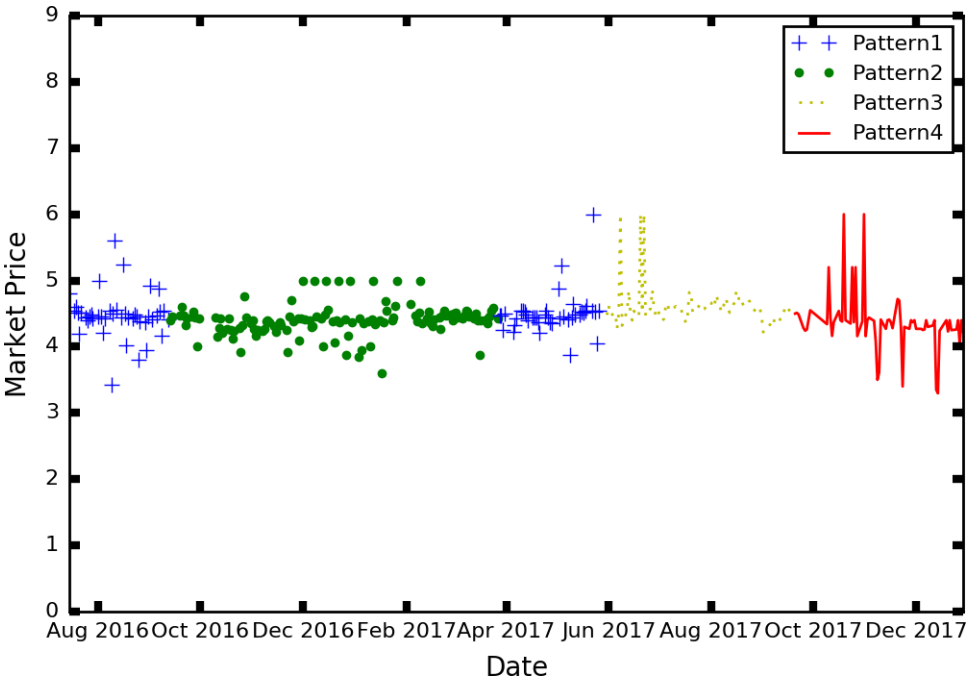
Next, we analyze the 4 clusters outputted by TICC to understand and interpret what state of the market they each refer to. Each cluster has a multilayer MRF network defining its structure. To analyze the result, we use complex network analysis to determine the relative "importance" of each node in the network. We plot the betweenness centrality score of each node in Table 5. We see that each of the 4 clusters has a unique "feature", with different factors having different points in each cluster. For example, the pattern1 has a non-zero score in only two of the four factors: Soybean Futures and Soybean-Oil Futures. As such, we would expect this cluster to refer to state in recession since the cash market volatility are volatile. Similarly, the scores of pattern2 are the most evenly, and it can be predicted that the soybean market is in a boom stage. We also see that the maximum and minimum values of cluster 3 and cluster 4 are quite different, so we can expect them to be in the transition state of pattern 1 and pattern 2, where cluster 3 is heavily affected by corn market and the soybean market is in a depression period. As such, a reasonable hypothesis according to the business cycle theory might be that the clusters refer to 1) recession, 2) boom, 3) depression, 4) recovery.

To validate that our experiment is interpretable, we now need to map the results back to the soybean market price to identify the accuracy of pattern. So we plot Figure3, coloring the timestamps according to their cluster assignments. Analyzing this graph, we empirically discover that each of the three clusters has their own characteristics and the recurring patterns do have some similarities. From this, we can perceive that our clustering is effective. Our experimental results support our claim that TICC

Table 5. Betweenness centrality for each factor in the four patterns. This score can be used as a proxy to show how "important" each factor is, and more specifically how much it directly affects the other factor values.

	Soybean Purchase	Corn Market	Soybean Futures	Soybean-Oil Futures
Pattern 1	0	0	5.833	10.33
Pattern 2	5.166	7.833	13.333	6.666
Pattern 3	6.233	37.198	1.833	13.733
Pattern 4	9.166	5.666	0	13.166

Figure 3. Schematic diagram of soybean market price in the same period



allows a time series of prices to be several typical patterns, and these patterns may recur over a certain period of time. The theoretical model fits the experimental well. Along this line of thinking, we can predict the direction of market prices by studying the transition probability of various models in the future.

CONCLUSION AND FUTUREWORK

In the context of this study, the empirical data showed that performance of the TICC methodology can discover soybean price pattern in a certain extent. We cluster each subsequence based on its correlation structure and define each cluster by a multilayer MRF to find various patterns in the data. Economic theory under each pattern has also been discussed. Furthermore, we mapped the results of the pattern discovery to the timeline of soybean market prices and found that our research was indeed explicable. Our study is an alternative approach in financial planning, as it allowed to detect and predict prices of agricultural products. Studying the pattern of agricultural products can enable us to grasp the price trend of market accurately, and then take effective measures to slow down the fluctuation of soybean market, which has important practical significance for stabilizing the production of agricultural products in China, increasing the output value of animal husbandry, and promoting the development of agricultural economy. In addition, pattern discover can be considered as one of the most important components in time series data mining systems. Using the a priori data

mining objective pattern, an optimal model will be chosen for short-term forecasting in the near future.

Under certain conditions and constraints, the paper only presents one repetitive pattern (Pattern 1) since we do not have much amount of price data. We believe when the data large enough to be used, we can receive more recurring pattern, even the relationship between the various patterns. This paper was in fact the first to apply the multi-time series clustering method to the field of agricultural product price analysis. Nevertheless, further works should be performed to learn how to accurately predict future prices based on Pattern Discovery. The properties that are discussed in this work, unquestionably, are receiving more and attention from other fields.

REFERENCES

Ahmad, M. H., Ping, P. Y., & Mahamed, N. (2014). Volatility Modelling and Forecasting of Malaysian Crude Palm Oil Prices. *Applied Mathematical Sciences*, *8*(124), 6159–6169. doi:10.12988/ams.2014.48650

Assis, K., Amran, A., & Remali, Y. (2010). Forecasting Cocoa Bean Prices Using Univariate Time Series Models. *Journal of Arts Science & Commerce*, *1*, 71–80.

Berndt, D. J., & Clifford, J. (1994). Using dynamic time warping to find patterns in time series. *Paper presented at the annual meeting of the Association for the Advance of Artificial Intelligence*, San Francisco, CA.

Boyd, S., Parikh, N., Chu, E., & Peleato, B. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, *3*(1), 1–122. doi:10.1561/2200000016

Catalao, J. P. S., Mariano, S. J. P. S., Mendes, V. M. F., & Ferreira, L. A. F. M. (2007). Short-term electricity prices forecasting in a competitive market: A neural network approach. *Electric Power Systems Research*, *77*(10), 1297–1304. doi:10.1016/j.epsr.2006.09.022

Edward, C. (1986). *Prescott*. Carnegie-Rochester Conference Series on Public Policy. Federal Reserve Bank of Minneapolis.

Ekananda, M., & Suryanto, T. (2018). MATEC Web of Conferences. Les Ulis, 150.

Ernst, J., Nau, J., & Bar-Joseph, Z. (2005). Clustering short time series gene expression data. *Bioinformatics (Oxford, England)*, *21*(1 Suppl 1), 159–168. doi:10.1093/bioinformatics/bti1022 PMID:15961453

Ferreira, L. N., & Zhao, L. (2016). Time series clustering via community detection in networks. *Information Sciences*, *326*(1), 227–242. doi:10.1016/j.ins.2015.07.046

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics (Oxford, England)*, *9*(3), 432–441. doi:10.1093/biostatistics/kxm045 PMID:18079126

Gionis, A., & Mannila, H. (2003). Finding recurrent sources in sequences. *Paper presented at the annual meeting of the International Conference on Research in Computational Molecular Biology*.

Golay, X., Kollias, S., Stoll, G., Meier, D., Valavanis, A., & Boesiger, P. (1998). A new correlation-based fuzzy logic clustering algorithm for fMRI. *Resonance Med*, *40*(2), 249–260. doi:10.1002/mrm.1910400211 PMID:9702707

Gray, R. M. (2001). Toeplitz and circulant matrices: A review. *Foundations and Trends in Communications and Information Theory*, *2*(3), 155–239. doi:10.1561/0100000006

Hallac, D., Vare, S., Boyd, S., & Leskovec, J. (2017). Toeplitz Inverse Covariance-Based Clustering of Multivariate Time Series Data. *Paper presented at the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Halifax, Canada. doi:10.1145/3097983.3098060

Harri, A., Nalley, L., & Hudson, D. (2009). The Relationship between Oil, Exchange Rates, and Commodity Prices. *Journal of Agricultural and Applied Economics*, *41*(2), 501–510. doi:10.1017/S1074070800002959

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Germany: Springer. doi:10.1007/978-0-387-84858-7

Kakizawa, Y., Shumway, R. H., & Taniguchi, N. (1998). Discrimination and clustering for multivariate time series. *Amer. Stat. Assoc*, *93*(441), 328–340. doi:10.1080/01621459.1998.10474114

Košmelj, K., Batagel, V. (1990). Cross-sectional approach for clustering time varying data. *Classification*, 799–109.

Liao, TW. (2005). Clustering of time series data—a survey. *Pattern discovery*, 38, 1857-1874.

Liu, Q. W. (2005). Price relations among hog, corn, and soybean meal futures. *Journal of Futures Markets*, *25*(5), 491–514. doi:10.1002/fut.20145

Nazlioglu, S., & Soytas, S. (2012). Oil price, agricultural commodity prices, and the dollar: A panel cointegration and causality analysis. *Energy Economics*, *34*(4), 1098–1104. doi:10.1016/j.eneco.2011.09.008

Parikh, N., & Boyd, S. (2013). Proximal algorithms. *Foundations and Trends in Optimization*, *1*(3), 127–239. doi:10.1561/2400000003

Peri, M., Baldi, L., & Vandone, D. (2013). Price discovery in commodity markets. *Applied Economics Letters*, *20*(4), 397–403. doi:10.1080/13504851.2012.709590

Pinheiro, C. A. O., & Senna, V. (2017). Multivariate analysis and neural networks application to price forecasting in the Brazilian agricultural market. *Ciência Rural*, *47*(1), 1–7. doi:10.1590/0103-8478cr20160077

Procacci, P. F., & Aste, T. (2018). Forecasting market states. SSRN Electronic Journal. doi:10.2139/ssrn.3215945

Ramirez, O. A. (2009). The Asymmetric Cycling of U.S. Soybeans and Brazilian Coffee Prices: An Opportunity for Improved Forecasting and Understanding of Price Behavior. *Journal of Agricultural and Applied Economics*, *41*(1), 253–270. doi:10.1017/S1074070800002674

Ramoni, M., Sebastiani, P., & Cohen, P. (2000). Multivariate clustering by dynamics. In *Proceedings of the 2000 National Conference on Artificial Intelligence*, San Francisco, CA (pp. 633-638).

Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *26*(1), 43–49. doi:10.1109/TASSP.1978.1163055

Smyth, P. (1997). Clustering sequences with hidden Markov models. In Advances in neural information processing systems (pp. 648-654).

Wang, C. (2016). Study on the influence factors of soybean price fluctuation in China. Dongbei university of finance and economics, Harbin.

Xiong, Y., & Yeung, DY. (2004). Time series clustering with ARMA mixtures. *Pattern discovery*, *37*, 1675-1689.

ENDNOTES

[1] Code and solver are available at https://github.com/davidhallac/TICC.

Hua Ling Deng, Department of Mathematics, Northeast Agricultural University, Professor, major in management science and engineering.

Yǔ Qiàn Sūn, Female, master's degree, major in management science and engineering. Second prize in "HUAWEI Cup" The 15th China Post-Graduate Mathematical Contest in Modeling.