Front. Comput. Sci.
DOI
RESEARCH ARTICLE

Use of sparse correlations for assessing financial markets

Xin Li, Guyu Hu, Yuhuan Zhou, Zhisong Pan（✉）
Army Engineering University of PLA, Nanjing 210007, China

Abstract The measurement of the partial correlation between stocks is of great significance for measuring the systematic risk of the stock market and the validity of a portfolio. The correlations between stocks usually change over time, especially in markets with high volatility (e.g., China's A share market). How to infer correlations with a strong temporal component, is a very important issue. This paper makes a relatively comprehensive correlation analysis of all the China's A-share stocks for the first time. Compared with conventional methods for estimating correlations between relatively few stocks with a large number of samples, this paper presents a method that infers the partial correlation graph between large numbers of stocks with relatively few samples based on Sparse Inverse Covariance Estimation (SICE). Analysis and experiments show that: (i) by adjusting the penalty coefficient, it is convenient to infer partial correlation graphs with different sparsity levels, which helps to find important information; (ii) through a series of experiments, we find that with an increasing penalty, the proportion of the companies in the same industry also gradually increase. In other words, the strongest correlations are often within the same industry. Meanwhile, when the market is relatively volatile, partial correlation graphs become very complicated and change over time.

Keywords Chinese stock market, partial correlation graph, sparse inverse covariance estimation.

1 Introduction

Stock correlation is a technical tool for studying the relationships between stocks, pairs trading, and industry classification. It is important for measuring the systemic risk in the stock market and the validity of portfolios. When constructing portfolio of assets [1,2], studying the linkage between stocks [8-10], or building the topologies of stock market [3,11,19], we all need to consider the correlation. However, correlations usually change over time, as the economy, policies, or state of the enterprises' operations change. It is therefore obvious that this change is more common within China's market, which is relatively volatile.

The Pearson correlation coefficient provides information about how similar the change is in terms of the price of a given pair of stocks. However, the correlation coefficient says nothing about whether a different stock will eventually control the observed relationship between the two stocks. A possible approach to overcome this issue is to make use of the statistical measure of partial correlation. Many literatures [3,20] have analyzed the difference between the two correlations, and the superiority of the partial correlation.

Suppose there are three stocks, A, B, and C, with a strong correlation between each other. If we recalculate the correlation between A and B after removing the effect of C, then we get the partial correlation between A and B. In recent years, the partial correlation in complex networks has become a very popular research topic, which has been applied to explain how a market index

Received Month dd, yyyy; accepted Month dd, yyyy

E-mail: xinyi2019@caa.columbia.edu

affects the relationships among stocks [5]. It is also used for industry classification when combined with clustering methods [6]. However, most of these studies focus on traditional statistical methods, solving the partial correlation of stocks with a large number of observation samples or a relatively small number of stocks [6,20,21]. For example, Jung [6] used ten-years of data to study the partial correlation of only 300 stocks. In doing this, the effects of the market index (factors) are being stripped away, not of all other stocks in the market. In a market with rapid style changes, how to find a relatively stronger correlation among stocks over a short period is a challenging and significant issue.

In this paper, we try to use three-years of data or even one-year's data to construct the partial correlation graphs for all the China's A-share stocks. We consider the inverse covariance (IC) matrix to construct the probability graph model. Because partial correlations correspond to the off-diagonal entries of the IC matrix of the data, estimating the partial correlations is usually achieved by the maximum likelihood estimation (MLE) of the IC matrix [7]. However, a limitation of MLE is that reliable estimates require the sample size of the data to be substantially larger than the number of stocks. Hence, we use graphical lasso, which also known as sparse inverse covariance estimation (SICE) [12], to model stock partial correlation graph. This method imposes a penalty on the MLE of the inverse covariance (IC) matrix, which allows us to obtain a reliable estimation with small sample sizes. Due to the monotonicity of SICE [13], we can obtain a sparse linkage corresponding to relatively stronger relationships among stocks. When we want to pick 100 of the (relatively) strongest edges from 1500 stocks, all we need to do is to raise the penalty appropriately. When we want to get more correlation information about the stocks, we only need to reduce the penalty, instead of setting the threshold manually to filter the correlation coefficients.

Specifically, we investigate the daily returns of all stocks in the Chinese stock market. We use the graphical lasso to construct a stock partial correlation graph for a period of time (such as three years or one year). Although a specific item of the sparse inverse covariance (SIC) matrix is not the true correlation efficient of the corresponding stocks because of the penalty imposed on the IC matrix, the structure and relative size of the SIC matrix are effective abstractions for the IC matrix. Based on the Jiang's work [18], we present a proof of the mathematical expression between the partial correlation coefficient and the entries of the SIC matrix in Appendix B. In the experiments we present, we try to answer the following questions. 1. In the market with faster style changes, how can we obtain effective partial correlations with small datasets over shorter time period, when the number of samples cannot meet the requirements of traditional statistical significance? 2. How will the SICE change when the market environment changes?

The remaining sections of this paper are organized as follows: Section 2 describes the SICE method. Section 3 reports on the data set used for the analysis and the experiments conducted as part of this study, while section 4 outlines the conclusions.

2 Method

We use an undirected graph to represent the partial correlation graph of the stock market. In the graph, a node stands for a stock and an edge stands for the partial correlation between stock A and stock B. The weights of the edges are estimated by the SICE.

Suppose we have P stocks and N observations (N ≪ P), i.e., $\{X_1, \ldots, X_p\}$. The inverse covariance (IC) matrix can be represented graphically. Two stocks are closely related if there is an edge between two nodes. The measurement for each stock is a log return $r_i(t)$, computed by:

$$r_i(t) = ln\big(Prc_i(t)\big) - ln\big(Prc_i(t-1)\big) \qquad (1)$$

where $Prc_i(t)$ denotes the adjusted closing price of stock i at time t. We assume the data follows a multivariate normality distribution with a mean μ and a covariance matrix Σ, $N(\mu, \Sigma)$, and the probability density function is given by:

$$p(x|\mu,\Sigma) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\} \quad (2)$$

Let $\Theta = \Sigma^{-1}$ denotes the IC matrix and assume that $\mu = 0$. First, we use the logarithmic likelihood function to solve with the above formula:

$$L(\Theta) = \log p(x_1, x_2, \ldots, x_n \,|\Theta)$$

$$= \sum_{i=1}^{n} \left\{ -\frac{p}{2}\log(2\pi) + \frac{1}{2}\log|\Theta| - \frac{1}{2} x_i^T \Theta x_i \right\}$$

$$= -\frac{np}{2}\log(2\pi) + \frac{n}{2}\log|\theta| - \frac{1}{2} tr\left(\Theta \sum_{i=1}^{n} x_i \, x_i^T \right) \quad (3)$$

If the $N \gg P$, we can obtain the $\widehat{\Theta} = S^{-1}$ by MLE, where $\widehat{\Theta}$ denotes the estimate of the IC matrix and S is the sample covariance matrix ($S = (1/n)\sum_{i=1}^{n} x_i \, x_i^T$). However, our hypothesis is $N \ll P$, and the S is irreversible. Hence, we introduce a sparse structure [16,17] and add the penalty term to the formula:

$$\widehat{\Theta} = argmax_{\Theta>0} log\big(det(\Theta)\big) - tr(S\Theta) - \lambda \parallel \Theta \parallel_1 \quad (4)$$

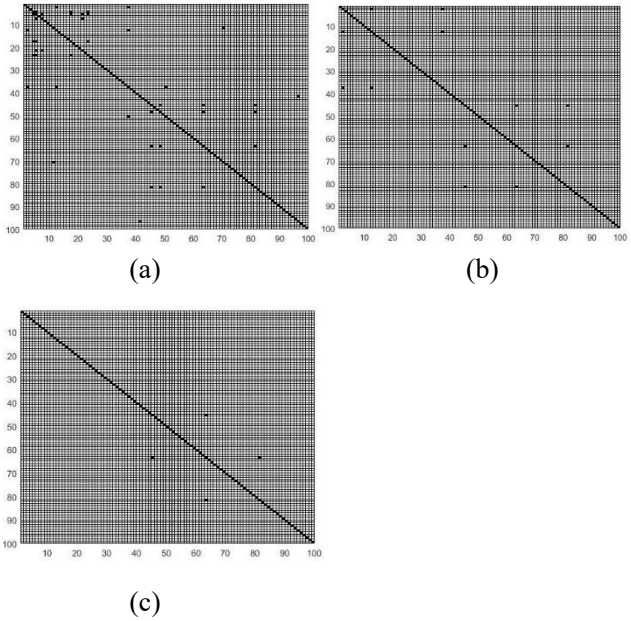
The formula above follows the same line as L_1-norm regularization [15]. $\det(\cdot)$ and $\mathrm{tr}(\cdot)$ denote the determinant and trace of a matrix respectively, S is the sample covariance matrix, $||\cdot||_1$ denotes the sum of absolute values of all the entries in a matrix, and $\lambda>0$ is the regularization parameter. We can see that the SICE aims to achieve a balance or tradeoff between the IC estimate and the sparsity. The λ controls the balance, where the larger the value of λ is, the sparser the Θ becomes. For this "Lasso" problem, we employ the algorithm in [12] to solve it in this paper (see the algorithm in the Appendix A).

The solved matrix $\widehat{\Theta}$ by SICE has a monotone property, i.e., if two nodes are not connected in the connectivity model for a certain λ, they will never become connected as λ increases [13]. When $\lambda=0$, it equals to IC matrix. If the matrix is invertible, we can obtain the relationships between the partial correlation and the entries in the IC matrix (see the proof in the Appendix B):

$$\rho_{X_i X_j * V\setminus\{X_i,X_j\}} = -\frac{\theta_{ij}}{\sqrt{\theta_{ii}\theta_{jj}}} \quad (5)$$

where $\Theta = (\theta_{ij})_{ij}$, θ_{ij} denotes the entry of IC matrix, and $V\setminus\{X_i, X_j\}$ represents the set excluding X_i and X_j. Therefore, $\rho_{X_i X_j * V\setminus\{X_i,X_j\}}$ is the partial correlation between X_i and X_j. As we can see from Eq. (5), the partial correlation coefficient is negatively correlated with the corresponding entry in the IC matrix.

With the monotone property, we know that if two stocks are connected (either directly or indirectly) at one level of sparseness ($\lambda = \lambda_2$), they will be connected at all lower levels of sparseness ($\lambda < \lambda_2$) [13]. This property can be used to identify how strong is partial correlation of each node (stock) X_k to its connectivity stocks. For example, let $S_k(\lambda_1)$ and $S_k(\lambda_2)$ be the sets of all the connectivity components of X_k with $\lambda_1 < \lambda_2$. Assuming that $S_k(\lambda_1) = \{X_i, X_j\}$ and $S_k(\lambda_2) = \{X_i\}$, this means that X_i is more strongly connected to X_k than X_j. Thus, by changing λ from small to large, we can obtain an order for the strength of the connection between pairs of stocks. As an example, Fig. 1 is an adjacency matrix drawn from the 100 stock data for the three-year period between 2012-2015, where $\lambda_{1\_a} < \lambda_{1\_b} < \lambda_{1\_c}$. When we can see clearly the corresponding set relationship, $S_k(\lambda_{1\_a}) \supseteq S_k(\lambda_{1\_b}) \supseteq S_k(\lambda_{1\_c})$.



(a) (b)

(c)

Fig.1: Graphs corresponding to the connection of the adjacency matrix for different values of λ corresponding to (a) λ_{1_a}, (b) λ_{1_b}, and (c) λ_{1_c}.

3 Data and Experiment

3.1 Datasets

For the long-term comparison, we take the daily closing price of the stocks of all China's A shares listed in 2006. After eliminating new shares that have been listed for less than half a year, there are remaining in total 1364 stocks. From January 2006 to January 2009, there are 729 observation samples used as the first phase of the data. Using the same method for the time period between

January 2009 to January 2012, there are 1572 stocks and 730 observation samples for the second phase of the data. The third phase of data is obtained from January 2012 to January 2015, with 2179 stocks and 726 observation samples. The fourth phase of data is obtained from January 2015 to January 2018, with 2516 stocks and 732 observation samples.

Considering short-term comparisons, we divide the fourth phase of data into three sample sets by years, that is from January 2015 to January 2016 (244 observations), January 2016 to January 2017 (244 observations), and January 2017 to January 2018 (244 observations). For stock quantity, in order to facilitate comparisons with long cycles, we still select 2515 stocks, excluding those in 2015 that had been listed for less than half a year.

The screening of the stock data includes two main aspects. In the long-term data, we exclude stocks suspended for more than 50 days for each phase of data. The number of stocks retained for 2006-2009, 2009-2012, 2012-2015 and 2015-2018 are therefore 787, 1357, 1553 and 1231, respectively. For the short-term data, we exclude stocks suspended for more than 20 days. The number of stocks retained in 2015, 2016 and 2017 are therefore 1335, 1825 and 1990, respectively. The historical stock data of this study was obtained from the Wind Financial Terminal, produced by Wind Information Inc. (the Wind Financial Terminal can be downloaded from http://www.wind.com.cn).

3.2 Long-term experiment

First, we experimented with the long-term stock data. In these experiments, we assigned values to λ from small to large. Because the sample size was smaller than the stock quantity, the penalty coefficient cannot be too small. Otherwise, divergent situation will happen. After a trial and error series of experiments, we find that when λ is less than 0.65 for some data, the model does not converge. Therefore, we chose $\lambda = 0.65$, $\lambda = 0.72$ and $\lambda = 0.81$ to represent the partial correlation graphs for these cases. The figures shown below are the sparse partial correlation graphs obtained by SICE where the node size is proportional to the node degree. Greater degree means the color is deeper and the node is bigger. With the enlargement of λ, we can see that the sequence of the connection diagram is becoming clearer. On the one hand, because of the monotony of SICE, we can adjust the size of λ to control the degree of sparseness to obtain the most relevant stock pair. On the other hand, due to the relatively short time, we can obtain timely and reliable correlations. For further research, such as pair trading, stock co-movement and so on, it can provide additional very valuable reference information.

3.2.1 When $\lambda = 0.65$

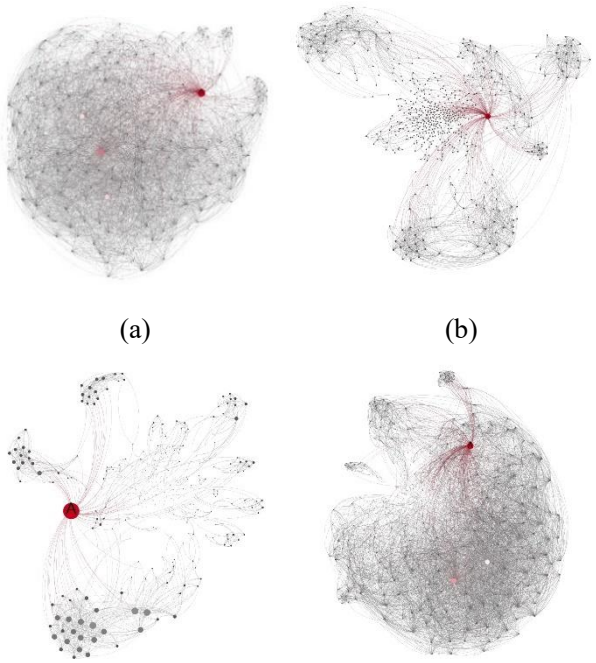


Fig.2: The connection between shares when λ is relatively small ($\lambda = 0.65$) for the periods (a) 2006 to 2009, (b) 2009 to 2012, (c) 2012 to 2015 and (d) 2015 to 2018.

When the penalty is relatively small, we can gain more correlations. Some are relativity stronger, but some weaker. From a structural point of view, we can see that the market index (factor) has a very wide impact on the entire market, with all stocks in the market directly or indirectly connected to it.

Fig. 2(a) shows the sparse solution to stocks from 2006 to 2009. We use the Shanghai Composite Index as the market index, denoted by the letter A. It is added to the original stock concentration with 788 nodes and 310,866 edges. After solving using the SICE algorithm, we get 485 nodes and 4,433 edges. From this figure, we

can see the node representing market index has the deepest color. This means there is a direct or indirect relationship between the market index and most individual stocks.

Fig. 2(b) presents the sparse partial correlation graph of stocks from 2009 to 2012. The market index is also represented by the letter A and again is added to the original stock concentration. There are a total of 1,358 nodes and 922,761 edges, while after solving by the SICE algorithm, we obtain 535 nodes and 1,960 edges.

Fig. 2(c) shows the stock from 2012 to 2015, where the original stock concentration has 1,553 nodes and 1,208,235 edges. After solving by the SICE algorithm, we get 389 sparse nodes and 1,071 linkage edges. Finally, Fig. 2(d) shows the stock from 2015 to 2018. The original stock concentration has 1,232 nodes and 759,528 edges. After solving using the SICE algorithm, the result is 713 nodes and 5,311 linkage edges.

By comparing the four figures above, we see that when the stock market volatility is greater, the partial correlation will be more complicated. As shown in Fig.2(a), from 2006 to 2009, China experienced a serious bull and bear markets. From January 2006 to October 2007, the market index (Shanghai Composite index) rose from 1,180 points to 6,124 points, but in the following year, it dropped back to 1,710 points by the beginning of November 2008. Similarly, from 2015 to 2018, due to the gradual liberalization of financial policies such as various margin financing, the stock allocation and other market activities increased exponentially. The Shanghai Composite index was 3,052.94 in February 2015 and less than half a year later, the highest point in June was 5,178.19. Then, the entire market began to control the leverage where the stock market experienced a cliff-like decline where in January 2016 it reached the lowest point of 2,638.30. However, in 2009-2012 and 2012-2015, it was relatively stable and during these two periods, the partial correlation is relatively clear.

3.2.1 When $\lambda = 0.72$

Due to the increase in λ, the model is sparser. The reserved nodes and edges indicate their stronger internal connections. We use the correlation size as a distance to analyze these nodes by k-means clustering. By including color labeling to indicate some of the major industries, we can see clearly that many of the shares in the same industry are connected (Fig. 3).

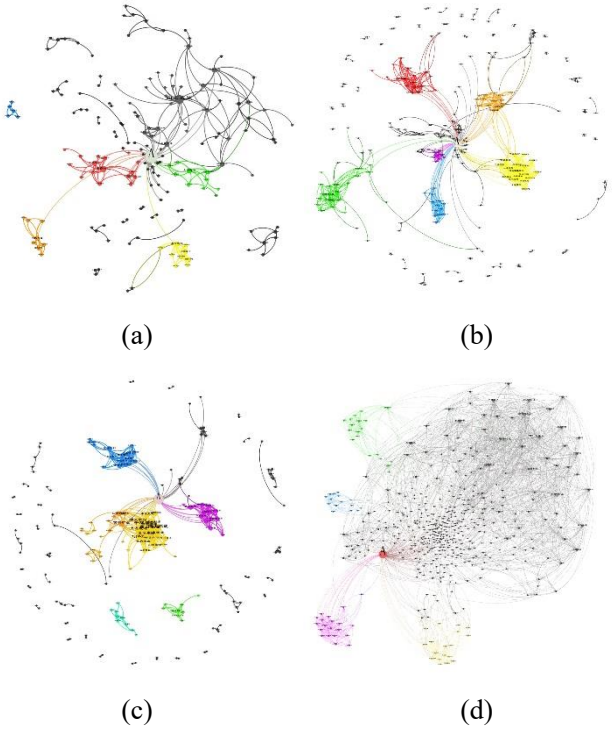


Fig. 3: The connection between shares when λ is increased ($\lambda = 0.72$) for the periods (a) 2006 to 2009, (b) 2009 to 2012, (c) 2012 to 2015 and (d) 2015 to 2018. Color coding is used to represented major industries.

In Fig. 3(a), there are 170 nodes and 254 edges. One can see that there is a lot of information that can be mined in the sub-connected graph, such as the sub-division of the various industries. Due to the large number of regions, we will not explain them one by one. We will show it from top to bottom, left to right. When we focus on the main business in the company, we find the agricultural stock is in the top left side indicated by blue. To the right in red are steel-industry stocks and right of this are the electricity stocks. The lower left shows the nonferrous metal industry in orange and the lower right part in yellow are the coal-industry stocks. In Fig. 3(b) there are 262 nodes and 705 edges. The red at the top-left are the steel industry stocks and to the right of these in orange are the nonferrous metal industry stocks. The purple segment in the middle is for the securities industry. The bottom-most green indicates the real estate industry, the blue to the right is the banking industry, and the lower-right yellow segments are the coal and energy industry.

In Fig. 3(c), there are 192 nodes and 473 edges. The blue part at the top is for the banking industry, with the yellow to the right being the coal and banking industry. The purple to the right indicates the securities industry. The dark green at the lower left are for the cement building industry and the lower-right green part indicates the real estate industry. In Fig. 3(d), there are 482 nodes and 1965 edges. The top green part represents the aeronautical and aeronautical industries. This part is electronic technology. The blue part below is the banking industry, with the lower left purple part the securities industry. On the right yellow color is the coal and energy industry. We also notice that there are many nonferrous metal and steel related stocks near this part. This corresponds to the strong resonance phenomenon caused by the policy of coal, nonferrous metals and steel in recent years.

When $\lambda = 0.65$, it is difficult to judge the effect of the stock concentration in the same industry, because the correlations are complicated. Here, we have initially seen that as λ increases, in the portions showing the stronger correlations, the majority of the shares are in the same industry. It also displays a gathering effect.

3.2.1 When $\lambda = 0.81$

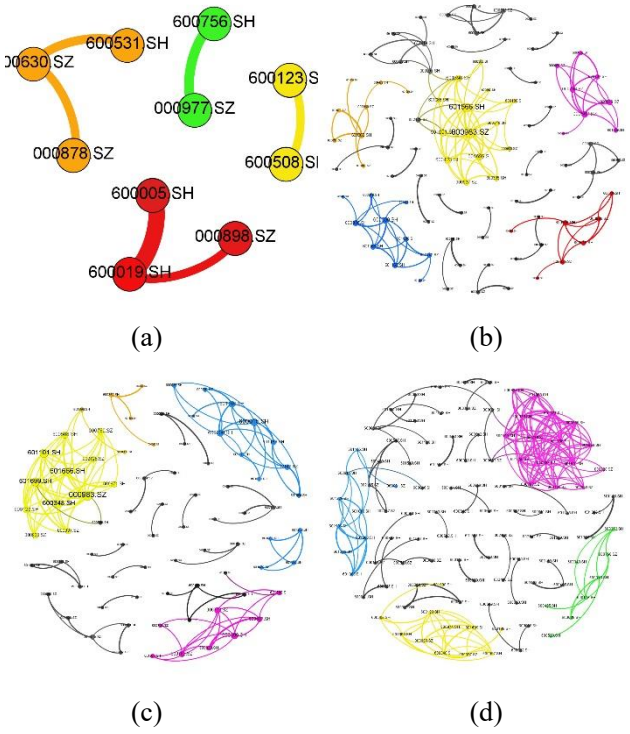


Fig.4: The case for when the value of λ is big enough, and the sparse partial correlation graph is clear enough, as well the sub-connected graph being sufficiently pure. (a) The time period from 2006 to 2009. (b) The time period from 2009 to 2012. (c) The time period from 2012 to 2015. (d) The time period from 2015 to 2018.

Due to the large coefficients now, there may be very few edges left in the dataset with fewer original nodes. However, according to the monotony of SICE, the correlation is also the strongest. We will compare the two cases when $\lambda = 0.65$ and $\lambda = 0.72$.

The colors for the different industries used in this figure are similar to those in Figure 3, namely green indicates the information industry, orange is for the nonferrous metal industry, yellow for coal and energy stock, purple stands for the securities industry, blue is for the banking industry, red is the steel industry, and green stands for the aviation and aerospace industry. In Fig. 4(a), there are 10 nodes and 6 edges which is very simple compared to the figures when the penalty $\lambda = 0.65$ and $\lambda = 0.72$. This indicates that most investors are new to the stock market. They may do not have a very clear understanding of the correlation between stocks. Therefore, the strongest correlations selected during this period are very few. In Fig. 4(b), there are 96 nodes and 141 edges. In Fig. 4(c), there are 78 nodes and 134 edges, while in Fig. 4(d), there are 106 nodes and 194 edges. We can see clearly that after partial correlation graphs becoming sparse enough, all stocks that are connected have a strong commonality. They either belong to the same industry or have a strong correlation for some special reasons. For example, in Fig. 4(a), for the information industry (green), they have the same controlling shareholder, Inspur Group Co. Ltd.

Meanwhile, in these four figures we can see that the market index (factor) no longer has a direct or indirect relationship with the majority of stocks as before. And, we find that stocks with edges connected are almost always in the same industry. This indicates that changes in the industrial environment have a significant impact on them and make these stocks have an internal linkage effect. For example, if a stock rises because of its own performance, this change will not connect with other stocks. But, if a stock rises because the industrial environment is good, then it will raise interest from similar stocks in the same industry, resulting in the linkage effect. In the stock market, which has limit fund, we can't demand that the industrial environment be good

and all stocks rise. What we observe more is that one of the connected stocks changes, with the others possibly undergoing co-movement in the future.

Through the experiments above, we can see the changes in the partial correlation graphs under different penalty sizes. We found the important impact factors, such as the Shanghai Composite Index. With the λ increasing, we can obtain the correlation between stocks more clearly and use them for arbitrage trading. Meanwhile, we find that the partial correlation graphs are clearer during the periods when the market is relatively stable. However, during two periods, 2006-2009 and 2015-2018, when the market is relatively volatile, the partial correlations are less clear. Therefore, we can conclude that the relationships are not static, but will change over time. When the market changes very sharply, the relationships will become more complicated. Therefore, it is not appropriate to establish a correlation graph of stocks for the volatile periods using a long period of data, such as ten years.

3.3 Industry comparison experiment

To make further comparisons of the data, we calculated the corresponding proportion of stocks in the same industry according the different values of λ under realistic industrial conditions, as shown in Table 1. We choose Wind classification (11 categories in total) and China Securities Regulatory Commission (CSRC) classification (19 categories in total) as the criteria for the industry divisions. The r represents the proportion of the same industry in a sub-connected graph. Suppose there are 100 nodes (stocks) in the connected graph of which 75 nodes belong to the same industry. Then the *r* equals 0.75. For example, from Table 1, considering the period 2006-2009 and the Wind classification criterion, when $\lambda = 0.65$, the proportion of $r = 1$ is 22%, and when we increase λ to 0.81, the proportion of $r = 1$ becomes 100%. For the latter, this means that all of the shares connected directly or indirectly are within the same industry, as Fig. 4(a) shows.

We can also see clearly that with increasing of λ, the proportion of the same industry increases significantly. For stocks that are not classified within the same industry, we find the following situations from further data analysis: (i) Two companies may be in the different industries, but these stocks have a strong conductivity, such as the cement and building materials industry; (ii) The controlling shareholders of the two companies overlap. For example, Jilin Aodong Co. Ltd. belongs to the pharmaceutical industry. However, among the top ten shareholders, there are nine belonging to securities funds, among which Guangfa Securities Co. Ltd. is the second largest shareholder with a shareholding of over 10%. The results also show that it has strong correlation with Guangfa Securities Co. Ltd.; (iii) Companies are seriously affected by common policies, such as the Chinese government's implementation of the supply-side reform and the restructuring of state-owned enterprises this year (2018). This will make the linkage phenomenon occur between enterprises with the same nature.

From the time perspective to compare the results, we can see that during the periods 2009-2012 and 2012-2015, when the market was relatively stable, the partial correlation graphs reflected more correlations between the same industry. However, during the other two periods, 2006-2009 and 2015-2018, we believe that the correlation of the stocks changed more quickly and in a more complicated manner. In other words, strong correlations in 2015 may not be correlated or weakly correlated in 2016.

Table 1 The corresponding proportions calculated according to the different values of λ under realistic industrial conditions. Note that the larger the value of *r*, the higher the proportion of the same industry.

		λ=0.65		λ=0.72		λ=0.81	
		Wind	CSRC	Wind	CSRC	Wind	CSRC
2006-2009	*r=1*	0.22	0.22	0.52	0.55	1.00	1.00
	0.75 ≤r<1	0.08	0.15	0.08	0.06	0.00	0.00
	0.5 ≤r<0.75	0.19	0.36	0.28	0.31	0.00	0.00
	r<0.5	0.51	0.27	0.12	0.07	0.00	0.00
2009-2012	*r=1*	0.56	0.47	0.72	0.67	0.89	0.88
	0.75 ≤r<1	0.15	0.19	0.13	0.12	0.02	0.01
	0.5 ≤r<0.75	0.21	0.22	0.10	0.15	0.07	0.10
	r<0.5	0.08	0.12	0.05	0.06	0.02	0.01
2012-2015	*r=1*	0.63	0.59	0.74	0.66	0.89	0.81
	0.75 ≤r<1	0.16	0.15	0.16	0.17	0.07	0.08
	0.5 ≤r<0.75	0.17	0.19	0.08	0.14	0.03	0.08
	r<0.5	0.04	0.07	0.02	0.03	0.01	0.03
2015-2018	*r=1*	0.25	0.28	0.39	0.42	0.91	0.82
	0.75 ≤r<1	0.13	0.17	0.12	0.12	0.03	0.10
	0.5 ≤r<0.75	0.28	0.37	0.26	0.31	0.05	0.07
	r<0.5	0.34	0.18	0.22	0.15	0.01	0.00

3.4 Long-term and short-term comparison experiment

We divided the data from 2015 to 2018 into three parts: January 1, 2015 to December 31, 2015, January 1, 2016 to December 31, 2016, and January 1, 2017 to December 31, 2017. The annual size of the observation sample is 244 trading days and the average number of nodes over the per year three years is about 1,600. To ensure the convergence of SICE, we will set λ to 0.85. After solving by SICE, there are 103 nodes and 134 edges in 2015, 114 nodes and 143 edges in 2016, and 15 nodes and 9 edges in 2017.

To ensure comparability, the number of edges should be as close as possible. Hence, we choose $\lambda = 0.81$ for the dataset, leading to 106 nodes and 194 edges. From Table 2, we can see that when the time period is shortened to one year, the sample quantity is greatly reduced compared to the longer-term experiment. However, the proportions of the same industry have not declined, but has risen.

According to the traditional method of seeking relevance, the unknown number is much more than the number of equations, hence, it is too difficult to solve. However, we can calculate the partial correlation matrix according to SICE. Although the absolute value does not represent the true value of correlation, the relative size can represent the strength of the correlation. The correlation represented by the preserved edges is stronger than the eliminated edges according to the monotonic properties of SICE. In comparison to the changes in the connections, we find that the partial correlation graph for 2015, 2016 and 2017 are different from each other.

Table 2 The corresponding proportions calculated for the time period of 2015-2018. Note that the proportion of *r* = 1is 91% by the Wind standard. By contrast, we use less data to get the proportion of *r* = 1 is 99% in 2015, 99% in 2016, and 100% in 2017.

	2015-2018		2015		2016		2017	
	Wind	CSRC	Wind	CSRC	Wind	CSRC	Wind	CSRC
r=1	0.91	0.82	0.99	0.99	0.99	0.99	1.00	1.00
0.75≤r<1	0.03	0.10	0.00	0.00	0.00	0.00	0.00	0.00
0.5≤r<0.75	0.05	0.07	0.01	0.01	0.01	0.01	0.00	0.00
r<0.5	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00

4 Conclusions

For the first time, this paper uses the SICE algorithm to construct a relatively comprehensive correlation map for all Chinese stocks. SICE was able to identify the sparse partial correlation graphs, and with the aid of its monotone property, it can also identify the order of inter-region connections in terms of the connection strength.

In the long-term experiments, we see how the correlation of the stock market is more complicated when the volatility is relatively violent. However, when the volatility is relatively smooth, the correlation of the stock market is relatively simple and clear. As λ increases, the proportion of the same industry also gradually increases. In other words, the strongest correlations are often in the same industry.

We also find that the correlations are dependent on the time period assessed, especially during more volatile periods. That is, in previous research, it was hysteretic or improper to study correlations using very long-term data, such as over ten years or more. To further understand this, we reduced the time period to one year. From the experimental results, we find that although the time period has been shortened, the ratio of the same industry is improved compared to the three-year cycle. At the same time, we see that in 2015, 2016 and 2017, the partial correlation graphs are different every year.

Acknowledgements This work was supported by the National Key Research Development Program of China [grant number 2017YFB0802800], The General Financial Grant from Jiangsu Natural Science Foundation for Youth in 2014 [grant number BK20140075]. And no potential conflicts of interests are reported by the authors.

Appendix A

The objective function can be written as:

$$Obj = \arg\min_{X>0}\{-\log\det X + \mathrm{trace}(SX) + \lambda\|X\|_1\} \quad (A1)$$

To get the sparse estimation of X, we should to derive the Eq. (A1):

$$W - S - \lambda\Gamma = 0 \quad (A2)$$

where the $W = X^{-1}$, and the Γ is a matrix of which the

entry γ_{ij} is:

$$\gamma_{ij} = \begin{cases} sign(W_{ij}), & W_{ij}^{-1} \neq 0 \\ [-1,1], & W_{ij}^{-1} = 0 \end{cases} \quad (A3)$$

We optimized the corresponding column of W in a block coordinate descent fashion:

$$W = \begin{pmatrix} W_{11} & w_{12} \\ w_{12}^T & w_{22} \end{pmatrix}, \quad S = \begin{pmatrix} S_{11} & s_{12} \\ s_{12}^T & s_{22} \end{pmatrix} \quad (A4)$$

Let $W = \Sigma$ and the $\Sigma^{-1} = \Theta = \begin{pmatrix} \Theta_{11} & \theta_{12} \\ \theta_{12}^T & \theta_{22} \end{pmatrix}$

Then $W \times \Theta = I$, we can get:

$$W = \begin{pmatrix} W_{11} & w_{12} \\ w_{12}^T & w_{22} \end{pmatrix} = \begin{pmatrix} \left(\Theta_{11} - \frac{\theta_{12}\theta_{21}}{\theta_{22}}\right)^{-1} & -W_{11}\frac{\theta_{12}}{\theta_{22}} \\ \cdots & \frac{1}{\theta_{22}} - \frac{\theta_{21}W_{11}\theta_{12}}{\theta_{22}^2} \end{pmatrix} \quad (A5)$$

Then:

$$w_{12} = -W_{11}\frac{\theta_{12}}{\theta_{22}} \quad (A6)$$

From the Eq. (A2), we know that:

$$w_{12} - s_{12} - \lambda\gamma_{12} = 0 \quad (A7)$$

Substituting Eq. (A6) into the Eq. (A7), then gives:

$$W_{11}\frac{\theta_{12}}{\theta_{22}} + s_{12} + \lambda\gamma_{12} = 0 \quad (A8)$$

Let $\beta = -\frac{\theta_{12}}{\theta_{22}}$, then gives:

$$W_{11}\beta - s_{12} - \lambda\gamma_{12} = 0 \quad (A9)$$

where $\gamma_{12} = \text{sign}(\theta_{12}) = \text{sign}\left(\frac{\theta_{12}}{\theta_{22}}\right) = -\text{sign}(\beta)$.

Solving the Eq. (A9) is equivalent to solving the problem:

$$\min\{1/2\beta^T W_{11}\beta - \beta^T s_{12} + \lambda\|\beta\|_1\}$$
$$= min\{\frac{1}{2}|W_{11}^{1/2}\beta - b|^2 + \lambda\|\beta\|_1\} \quad (A10)$$

where $b = W_{11}^{-1/2}s_{12}$. Thus, we transform the problem of graphical lasso to the problem of lasso.

$$\hat{w}_{12} = W_{11}\hat{\beta}, \quad \hat{w}_{12} = s_{22} + \lambda$$
$$\hat{\theta}_{12} = \frac{\hat{\beta}}{\hat{w}_{22} - \hat{\beta}\hat{w}_{12}}, \quad \hat{\theta}_{22} = \frac{1}{\hat{w}_{22} - \hat{\beta}\hat{w}_{12}}$$

For each iteration, we rearrange the rows and columns of W so that the target column is in the last.

Appendix B

Let $x_1, x_2, \cdots, x_k$ be k variables, and $\rho_{x_1 x_2 * v\setminus\{x_1,x_2\}}$ be the partial correlation between x_1 and x_2 given all the others' impacts. Then, x_1 and x_2 can be written as:

$$\widehat{x_1} = \eta_{10} + \eta_{13}x_3 + \cdots + \eta_{1k}x_k \quad (B1)$$

$$\widehat{x_2} = \eta_{20} + \eta_{23}x_3 + \cdots + \eta_{2k}x_k \quad (B2)$$

To simplify this calculation, all the variables $(x_1, x_2, \ldots, x_k)$ have been centralized, that is $E(X) = 0$. Now, let

$$X_{12} \triangleq (X_3, X_4, \cdots, X_k) \quad (B3)$$

$$\Gamma_j \triangleq (\eta_{j3}, \eta_{j4}, \cdots, \eta_{jk})^T, \quad j = 1,2 \quad (B4)$$

Then

$$X_j = X_{12}\Gamma_j + e_j, \quad j = 1,2 \quad (B5)$$

where e_j is defined as the residuals of the regression equation. Minimizing the sum of the squared residuals requires us to choose Γ_j such that

$$\min S(\Gamma_j) = e_j^T e_j = (X_j - X_{12}\Gamma_j)^T (X_j - X_{12}\Gamma_j) \quad (B6)$$

The necessary condition for a minimum Eq. (B6) can get

$$X_{12}^T \cdot X_{12} \cdot \Gamma_j = X_{12}^T \cdot X_j, \quad j = 1,2 \quad (B7)$$

Denoting now

$$\Lambda = \begin{pmatrix} \| x_3 \| & & 0 \\ & \ddots & \\ 0 & & \| x_k \| \end{pmatrix} \quad (B8)$$

Then

$$\Lambda^{-1}X_{12}^T X_{12}\Lambda = (\rho_{ij})_{(k-2)\times(k-2)} \triangleq R_{12} \quad (B9)$$

$$\Lambda^{-1} \cdot X_{12}^T \cdot X_j = \| X_j \| \cdot (\rho_{ij})_{(k-2)\times 1} \triangleq \| X_j \| \cdot R_j \quad (B10)$$

Substituting Eq. (B9) and Eq. (B10) into the Eq. (B7) then gives

$$R_{12}\Lambda\Gamma_j = \| X_j \| R_j \quad (B11)$$

$$\Rightarrow \Gamma_j = \| X_j \| \Lambda R_{12} R_j \tag{B12}$$

With the residual:

$$e_j = X_j - \hat{X}_j = X_j - X_{12}\Gamma_j \tag{B13}$$

Then

$$\left\|e_j\right\|^2 = e_j^T \cdot e_j = \left\|X_j\right\|^2 - \Gamma_j^T X_{12}^T X_j$$

$$= \left\|X_j\right\|^2 \left(1 - R_j^T R_{12}^{-1} R_j\right), \ j = 1,2 \tag{B14}$$

$$e_1^T \cdot e_2 = X_1^T X_2 - X_1^T X_{12}\Gamma_2 - \Gamma_1^T X_{12}^T X_2 + \Gamma_1^T X_{12}^T X_{12}\Gamma_2$$
$$= X_1^T X_2 - \| X_1 \| \cdot \| X_2 \| R_1^T R_{12} R_2 \tag{B15}$$

Let, $R \triangleq | (\rho_{ij})_{k\times k} |$ be a correlation matrix, and a_{ij} be the cofactors of ρ_{ij}, then

$$R = \begin{vmatrix} \rho_{11} & \rho_{12} & R_1^T \\ \rho_{21} & \rho_{22} & R_2^T \\ R_1 & R_2 & R_{12} \end{vmatrix} \tag{B16}$$

In the above equation (A16), R is a symmetric matrix and $\rho_{11} = \rho_{22} = 1$, then

$$a_{11} = \begin{vmatrix} \rho_{22} & R_2^T \\ R_2 & R_{12} \end{vmatrix} = | R_{12} | (1 - R_2^T R_{12}^{-1} R_2) \tag{B17}$$

$$a_{22} = \begin{vmatrix} \rho_{11} & R_1^T \\ R_1 & R_{12} \end{vmatrix} = | R_{12} | (1 - R_1^T R_{12}^{-1} R_1) \tag{B18}$$

$$a_{12} = -\begin{vmatrix} \rho_{21} & R_2^T \\ R_1 & R_{12} \end{vmatrix} = -| R_{12} | (\rho_{21} - R_2^T R_{12}^{-1} R_1) \tag{B19}$$

Therefore

$$\rho_{x_1 x_2 * v\backslash\{x_1,x_2\}} \triangleq \frac{e_1^T e_2}{\| e_1 \| \cdot \| e_2 \|}$$
$$= \frac{\rho_{21} - R_2^T R_{12}^{-1} R_1}{\sqrt{(1 - R_1^T R_{12}^{-1} R_1)(1 - R_2^T R_{12}^{-1} R_2)}}$$
$$= -\frac{a_{12}}{\sqrt{a_{11} a_{22}}} \tag{B20}$$

Now, let $\mathrm{R}^\star$ be the adjoint matrix of R

$$\mathrm{R}^\star = \begin{pmatrix} a_{11} & \cdots & a_{11} \\ \vdots & \ddots & \vdots \\ a_{1n} & \cdots & a_{nn} \end{pmatrix} \tag{B21}$$

Then

$$\mathrm{R}^{-1} = \frac{\mathrm{R}^\star}{\det(\mathrm{R})} = \begin{pmatrix} \frac{a_{11}}{det(R)} & \cdots & \frac{a_{n1}}{det(R)} \\ \vdots & \ddots & \vdots \\ \frac{a_{1n}}{det(R)} & \cdots & \frac{a_{nn}}{det(R)} \end{pmatrix}$$
$$= \begin{pmatrix} \theta_{11} & \cdots & \theta_{1n} \\ \vdots & \ddots & \vdots \\ \theta_{n1} & \cdots & \theta_{nn} \end{pmatrix} \tag{B22}$$

which then results in

$$\rho_{x_1 x_2 * v\backslash\{x_1,x_2\}} = -\frac{a_{12}}{\sqrt{a_{11} a_{22}}} = -\frac{\theta_{12}}{\sqrt{\theta_{11}\theta_{22}}}$$

From the above deduction, we can easily extend the case to $\mathrm{x_i}$ and x_j

Finally: $\rho_{x_i x_j * v\backslash\{x_i,x_j\}} = -\frac{\theta_{ij}}{\sqrt{\theta_{ii}\theta_{jj}}}$

References

1. Markowitz H. Portfolio Selection[J]. Journal of Finance, 1952, 7(1):77-91.

2. Ang A, Chen J. Asymmetric correlations of equity portfolios [J]. Social Science Electronic Publishing, 2002, 63(3):443-494.

3. Kenett D Y, Tumminello M, Madi A, et al. Dominating Clasp of the Financial Sector Revealed by Partial Correlation Analysis of the Stock Market[J]. Plos One, 2010, 5(12):e15032.

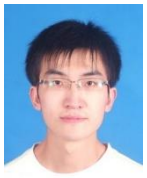
4. Longin F, Solnik B. Is the correlation in international equity returns constant: 1960–1990? [J]. Cepr Financial Markets Paper, 1995, 14(1):3-26.

5. Shapira Y, Kenett D Y, Ben-Jacob E. The Index cohesive effect on stock market correlations[J]. European Physical Journal B, 2009, 72(4):657.

6. Jung S S, Chang W. Clustering stocks using partial correlation coefficients[J]. Physica A Statistical Mechanics & Its Applications, 2016, 462:410-420.

7. Wiley. On the Inverse of the Covariance Matrix in Portfolio Analysis[J]. Journal of Finance, 1998, 53(5):1821-1827.

8. Eun C S, Shim S. International Transmission of Stock Market Movements[J]. Journal of Financial & Quantitative Analysis, 1989, 24(2):241-256.
9. Chiang T C, Bang N J, Li H. Dynamic correlation analysis of financial contagion: Evidence from Asian markets[J]. Journal of International Money & Finance, 2007, 26(7):1206-1228.
10. Forbes K, Rigobon R. Only Interdependence: Measuring Stock Market Co-movements[J]. 2002, 42.
11. Tumminello M, Lillo F, Mantegna R N. Correlation, hierarchies, and networks in financial markets[J]. Journal of Economic Behavior & Organization, 2010, 75(1):40-58.
12. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso[J]. Biostatistics, 2008, 9(3):432-441.
13. Huang S, Li J, Sun L, et al. Learning brain connectivity of Alzheimer's disease by sparse inverse covariance estimation[J]. NeuroImage, 2010, 50(3): 935-949.
14. Efron B, Hastie T, Johnstone I M, et al. Least angle regression[J]. Annals of Statistics, 2004, 32(2): 407-499.
15. Tibshirani R. Regression shrinkage and selection via the lasso: a retrospective[J]. Journal of The Royal Statistical Society Series B-statistical Methodology, 2011, 73(3): 273-282.
16. Banerjee O, Ghaoui L E, Daspremont A, et al. Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data[J]. Journal of Machine Learning Research, 2008: 485-516.
17. Yuan M, Lin Y. Model selection and estimation in the Gaussian graphical model[J]. Biometrika, 2007, 94(1): 19-35.
18. Jianghong Ma. Some questions about multiple linear regression equations [J]. Journal of Xi ' an highway transportation university, 1994(1):89-94.
19. Xue G, Hu Z, Tianhai T, et al. Development of stock correlation networks using mutual information and financial big data[J]. PLOS ONE, 2018, 13(4):e0195941-.
20. Wang G J, Xie C, Stanley H E. Correlation structure and evolution of world stock markets: Evidence from Pearson and partial correlation-based networks[J]. Computational Economics, 2018, 51(3): 607-635.
21. Sai K H S, Pal M, Manimaran P. Multifractal detrended partial cross-correlation analysis on Asian markets[J]. Physica A: Statistical Mechanics and its Applications, 2019: 121778.



Li Xin, Ph.D., Department of Computer Science and Technology, Army Engineering University of the PLA. In 2012, he went to Guanghua School of Management, Peking University to exchange studies. Main research areas: machine learning, financial data analysis, graph models, etc.

Zhisong Pan received the Diploma degree in computer science and technology from the PLA Information Engineering University, Zhengzhou, China, in 1996 and the Ph.D. degree in computer science and technology from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2003. He is currently a Professor with the Army Engineering University of PLA, Nanjing. His current research interests include deep learning, machine learning, and pattern recognition.

Guyu Hu, Professor and doctoral tutor of the Institute of Command Automation of Army Engineering University of the PLA. He is a senior member of the Chinese Institute of Electronics, an academic member of the Nanjing Computer Forum of the Chinese Computer Society Youth Computer Technology Forum (YOCSEF), and a member of the first communication software technology committee of the China Communications Society.

Yuhuan Zhou, Ph.D., Department of Computer Science and Technology, Army Engineering University of the PLA. Main research areas: deep learning, speech recognition, pattern recognition, data stream recognition, etc.