From correlations to eigenvalues: mining information from financial data using random matrix theory

Yong Tang[a,d,*], Jason Jie Xiong[b], Zi-Yang Jia[c], Yi-Cheng Zhang[d]

[a] *School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, 610054, China*
[b] *Department of Computer Information Systems and Supply Chain Management, Walker College of Business, Appalachian State University, Boone, NC 28608, USA*
[c] *Department of Computer Science, Rutgers University, Piscataway, NJ 08854, USA*
[d] *Department of Physics, University of Fribourg, Chemin du Musée 3, CH-1700 Fribourg, Switzerland*

Abstract

Stock markets are typical complex systems that contain embedded information. Data mining and knowledge discovery methods are needed to extract those hidden information from noises. This research systematically analyzes the behaviors of correlations among stock prices and the eigenvalues for correlation matrices by utilizing Random Matrix Theory (RMT) for Chinese and US stock markets. Results suggest that most eigenvalues of both markets fall within the predicted distribution intervals by RMT, while some larger eigenvalues fall beyond the noises and carry market information. The largest eigenvalue is found representing the market and a good indicator for averaged correlations. Further, the averaged largest eigenvalue shows similar movement with the index for both markets. The analysis demonstrates the fraction of eigenvalues fall beyond the predicted interval pinpoint major market switching points. It is identified that the average of eigenvector components corresponding to the largest eigenvalue switch with market itself. The investigation on the second largest eigenvalue and its eigenvector suggests that Chinese market is dominated by four industries while the US market contains three leading industries. The study later investigates how it changes before and after a market crash, revealing that two markets

*Corresponding author
Email address: tangyong@uestc.edu.cn (Yong Tang)

behave differently, in which a major market structure change is observed in the Chinese market but not in the US market. The research contributes to the literature by providing a systematical studies from random matrix theory modelling approaches for the two major stock markets. The results shed new lights on mining hidden information from stock market data.

Keywords: Financial big data, Market structure, Random matrix theory, Eigenvalue analysis

1. Introduction

Thanks to the availability of financial data in a wide range of frequencies from tick to daily, it's possible to apply data mining and knowledge discovery methods beyond traditional finance but from data science, network analysis, and even physics, etc. The asset prices in the markets are results of a complicated dynamics of spreading and reacting of market signals and information. The market structures are embedded in the prices movements which are normally correlated with each other. As a starting point for the underlying cornerstones of finance theories like Modern Portfolio Theory (MPT) [**?**] and Capital Asset Pricing Model [**?**], the correlation information of assets prices is always at the heart for theoretical researches and finance industrial practices in portfolio management and risk management etc.

For a portfolio of N stocks, we need a correlation matrix with $N \times N$ elements to describe the pairwise relationships. With the increase of N, the number of possible relationships grows rapidly making it difficult and challenging to calculate or analyze. To extract the hidden structure and essential information, it's necessary to simplify the network by filtering the less important elements in order to make it still feasible to analyze portfolios even with a very large N. In the past few years, we see some methods have been introduced to simplify the stock matrices. To study the correlation behaviors of the financial markets, correlation matrix is constructed from the price time series before we apply methods and techniques such as *Principal Component Analysis* (PCA) [**?** **?**

? **?**], *Multidimensional Scaling* (MDS) [**?**], *Factor Analysis* (FA) [**?**], *Minimum Spanning Tree* (MST) [**?** **?**], *Hierarchical Clustering* (HC) [**?** **?** **?**], and *Singular Value Decomposition* (SVD) [**?**]. In this study, we discuss the use of *Random Matrix Theory* (RMT) in the study of financial markets. On the other side of correlation matrix simplifying lies the validation, which statistically validate the matrix and keep those validated elements, in thus to achieve a simple matrix with less noises and easy for analysis.

The validations provide statistical confidences of the results or insights extracted from the validated matrices. The underlying idea to design a validation is to compare the empirical matrices with random ones generated from same distributions, random shuffles, or statistical tests with which the null hypothesis is setup to be tested with empirical data. Any deviations from these *benchmark* are considered as noises and should be filtered. In the similar manner, given an empirical correlations matrix (and the derived distance matrices for the networks), we can consider a random matrix with the same size. A null hypothesis can be introduced to test the statistical validation of each elements of the original empirical matrix by comparing the distributions. The basic idea is that any deviations from the random distribution is believed as validated with genuine information of the system, while those fall within the random distribution are pure random noises contains no system information. Random Matrix Theory (RMT), originally been applied to study the nuclear activities [**?**] in the early 1960s, is first introduced into the study of financial markets by [**?**] and recent years see new advances in applying RMT in finance studies and applications [**?** **?** **?**].

The Chinese and US markets are two leading markets but different in maturity and regulation, in this study, based on a dataset covering 9 years of stock prices, we systematically investigate two stock markets of China and US using RMT to study and compare the correlation properties and the dynamics of eigenvalues as well as eigenvectors. The findings revealed that the two stock markets are both similar and different in many ways. The results add new insights on market behaviors with implications for finance applications such as

portfolio management and optimization, market risk monitoring, trading strategy design. Meanwhile, this study also serves as a framework for data mining and knowledge in financial big data using RMT.

This work is organized as follows. First, we review the literatures in Section 2. Then in Section 3 we describe the dataset of two markets and methods used in this study. The properties of correlation matrices are discussed in Section 4. Using RMT, in Section 5, the properties and behaviors of eigenvalues and eigenvectors are analyzed with an investigation of a market switch study. Finally, Section 6 presents conclusions and discussions.

2. Literature review

2.1. Random matrix theory

The eigenvalue distribution of a pure random matrix C_{random} with the same size of C follows

$$p(\lambda)_{random} = \frac{Q}{2\pi} \frac{\sqrt{(\lambda_{\max} - \lambda)\,(\lambda - \lambda_{\min})}}{\lambda}, \tag{1}$$

where λ_{min} and λ_{max} are the theoretical minimum and maximum eigenvalue bounds of random matrix, the Q is the ratio of L/N satisfying the requirement that $Q > 1$, $L \to \infty$, and $N \to \infty$. Using the empirical data, we can also get the empirical distribution as

$$p(\lambda)_{random} = \frac{1}{N} \frac{dn(\lambda)}{d\lambda}. \tag{2}$$

Theoretically, with the knowledge of Q, we can determine the theoretical eigenvalue bounds as

$$\lambda_{min,max} = 1 + \frac{1}{Q} \mp 2\sqrt{\frac{1}{Q}}. \tag{3}$$

With these calculations, we can construct and determine the theoretical distribution of a null hypothesis random matrix. The empirical eigenvalues fall within the interval of $[\lambda_{min}, \lambda_{max}]$ are pure random noises, and those fall beyond the interval are the validated eigenvalues carrying true information of the system. In this way, we also get the validated corresponding eigenvectors for

those validated eigenvalues. Also, we can go further to investigate the statistical validation of the eigenvectors. The distribution of the eigenvector components in v_i for eigenvalue λ_i follows the *Porter-Thomas distribution* [**?**] as

$$P(v_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{v_i^2}{2}}, \tag{4}$$

with which we can validate the eigenvector components by comparing the distributions. [**?**] report that the distribution of eigenvector components of the largest eigenvalues shows great difference from the theoretical predictions.

2.2. Random matrix theory approach in financial markets

Recently, there is a emergence of researches using RMT in financial markets to filter noises and reveal embedded market properties. The cross-correlations of stock prices are studied using RMT to identify correlated relationships [**?**]. Furthermore, free random variables are applied in RMT analysis in financial time series [**?**]. RMT has also be applied to return estimation and asset allocation in Markowitz mean-variance optimization [**?**]. Using time shifted series, the lagged correlation matrices are studied from RMT approach to compute eigenvalue density and identify deviations [**?**].

For an empirical correlation matrix C of size $N \times N$ generated from N returns series of length L, we can construct the elements as,

$$C = \frac{1}{L} M M^T, \tag{5}$$

where M is a $N \times L$ matrix with normalized return $y_i(t)$ for each stock at every time t, where

$$y_i(t) = \frac{Y_i(t) - \langle Y_i(t) \rangle}{\sigma_i}. \tag{6}$$

Then we can further calculate the Markowitz's *efficient frontier curve* [**?**] for the portfolio which satisfies

$$\frac{\partial}{\partial w_i} (R_w - \alpha Y_w) \big|_{w_i = w_i^*} = 0, \tag{7}$$

where R_w is the mean risk of the portfolio, and α is a linear coefficient, w_i and w_i^* stands for the actual weights and the optimized weights [**?**], respectively.

Considering that for a correlated portfolio, the mean risk can be denoted as

$$R_w = \sum_{i,j=1}^{N} w_i w_j C_{ij}. \tag{8}$$

This marries the correlation matrix with the portfolio risk/return. So, once we construct the correlation matrix using Eq. 5, it can be used in calculating the efficient frontier. Now, the only concern is the validation of this correlation matrix.

Since the introduction of RMT into financial market study, many literatures investigated different markets. [**?**] also point out that the lower bound is positive and no eigenvalues fall between 0 and λ_{min} also vanish above λ_{max}. Since the empirical values of N and L are limited far from ∞, the edges are blurred with some eigenvalues fall beyond the bounds [**?**]. The distribution of the spacings of eigenvalues $s \equiv \lambda_{i+1} - \lambda_i$ are found to agree with *Wigner distribution* of the energy spacing levels [**?**]. This provides evidences indicating that the empirical correlation matrix is consistent with its random matrix counterpart. Many empirical studies reveal that only a small faction of eigenvalues and their corresponding eigenvectors contain system information while most are embedded in noises. It's reported that the portion of largest eigenvalues deviating from the theoretical prediction of the counterpart random matrix is 6% in [**?**], 4.7% in [**?**], 2% in [**?**], 11% in [**?**], and 1% in [**?**]. It's also found that the average of correlations in the correlation matrix can be well estimated from the largest eigenvalue as

$$\lambda_{max}/N \sim \langle C_{ij} \rangle \,. \tag{9}$$

Using the Shanghai Interbank Offered Rate (SHIBOR) daily quotes of 16 components commercial banks from October 8, 2006 to August 31, 2012 covering 1457 trading days, it's verified that the largest eigenvalue λ_{max} is a good estimator of the average correlation of the correlation matrices constructed from a sliding window approach [**?**]. The same results are also reported in [**?** **?**] revealing that the average correlation co-moves with the largest eigenvalue for

the component stocks of S&P500. The absolute value of scalar product

$$S_{ij} = |v_i \cdot v_j| \tag{10}$$

is used to describe the similarity of two eigenvectors [**?**], from the meanings of scalar product, we see that the similarity S_{ij} is determined by the cos of the angle between v_i and v_j [**?**]. For normalized eigenvectors, the value of S_{ij} ranges from 0 to 1, in other words, the two eigenvectors change from orthogonal to exact the same. [**?**] report that the effect of noises on the risk gets insignificant in measure of fixed portfolio while remain important for optimized portfolio for small values of N/L. This indicates that the correlation matrix can still be valid in traditional risk management and portfolio optimization even most information is covered by noises. By using simulating methods, many correlation matrix filtering approaches are tested and the approaches based on random matrix theory are found perform consistent well in all cases [**?**]. The eigenvalue distribution of emerging stock market is reported to be different from developed markets though correlation distributions and other properties are similar. Methods based on clustering for portfolio optimization and the effective size determination are proposed and the results are found to be improved compared to RMT approaches [**?**], which indicates that RMT might be further combined with other methods in filtering matrix and optimizing a portfolio. Following the RMT approach, the largest eigenvalues are found to be responsible for the market mode, by removing this, the correlation matrix is cleaned to reveal the topological structures [**?**]. Taking into account the signs of eigenvector components, sub-sectors of positive and negative signs can be derived from sectors in anti-correlation. The sub-sectors are detected with strong appearances in Chinese stock market but weaker in US stock market [**?**]. US and British stock exchanges are studied by using RMT on the asymmetric correlation matrix with a lag of τ [**?**]. The details of the residual noise part for a market are studied revealing that the noise band is composed of more sub-bands [**?**]. Using RMT, the Chinese stock market is studied [**?**], a similar anti-correlation relationship between sub-sectors are studied [**?**], the results show that the prominent sector

structure exists and the distribution of eigenvalues also reveal that the market is likely to be influenced by the global financial crisis and policies applied by the Chinese government. In a further study on the sub-sectors of a stock market, local interaction structures are found change during financial crises [**?**]. The sign information of components in eigenvectors is again used to detect the sub-sector anti-correlations [**?**]. Focus on how the credit market and stock market behave before and after a financial crisis, RMT is applied and find that the largest eigenvalue of credit market precedes that of stock market [**?**], this indicates that the pattern changes of eigenvalues have potential implications in the understanding of inter relationships between different markets.

To decompose the original correlation matrix into different eigenvalue modes, elements are weighted as

$$C_{ij} = \sum_{\alpha=1}^{N} \lambda_i C_{ij}^{\alpha}, \tag{11}$$

where $C_{ij}^{\alpha} = u_i^{\alpha} u_j^{\alpha}$ stands for the correlation element for αth eigenvalue [**?**]. The study of [**?**] provides a study of the eigenvalues spectrum for Chinese stock market in a sliding window approach. The *inverse participation ratio* is defined as

$$I^k = \sum_{l=1}^{N} \left[u_l^k\right]^4, \tag{12}$$

where u_l^k is the components of eigenvector v^k, to measure the deviation degree of eigenvectors [**?**]. A criterion of *fractional Gaussian noise* (fGn) is used to evaluate the autocorrelation matrix of stocks showing agreement with fGn though the stock returns are non-Gaussian [**?**]. The study of [**?**] adds new evidence that not all eigenvalues fall into the theoretical interval predicted by the random matrix are pure random noise but still carry some information.

3. Data and correlation matrices

3.1. Data

In this paper, we study the stock markets of China and United States, the former is a typical representative of emerging countries with fast growing GDP

rate and influence on global economies, while the latter is the most established and developed economy in the world. In order to study the major stocks of each market, we focus on the component stocks of the major indices of the two stock markets: China Securities Index 300 (CSI300) for the Chinese stock market and Standard & Poor's 500 (S&P500) for the US market. In our study, we cover a period of 9 years starts on 04/01/2007 and ends on 06/11/2015 with 2149 trading days for CSI300 and 2228 trading days for S&P500. The reason why the two markets have different numbers of trading dates is that the two markets have different trading calendars. Index and all component stocks daily price data of CSI300 are retrieved from the CSMAR Solution Database of Shenzhen GTA Education Tech. Ltd. We download the S&P500 index and component stocks daily prices data through Yahoo Finance service. Since not all stocks are traded on each trading date, so we only select those CSI300 stocks with at least 2000 trading dates and without continuous 100 non-trading dates, this selection results a final set of 163 stocks. For S&P500, we select those stocks with at least 2100 trading dates, and in results we get 468 stocks. After stocks selection, we take the prices on the available closest trading date to fill the non-trading dates.

3.2. Correlation matrices

In a financial market, for example a stock market, the prices of stocks fluctuate constantly showing complicated behaviors. It's important to investigate the performance of individual stocks as well as the interactive behaviors among stocks. To evaluate the interactive co-movement behaviors among the prices of asserts, the correlation is a fundamental concept widely used in studies of price dynamics and also are used in the traditional theories. When correlation is considered, in traditional theories, like in MPT where the correlation matrices are actually inputs for the portfolio optimization [**?**], the correlation is assumed as fixed, but in real world, the correlations are fluctuating and demonstrate some collective behaviors in market crashes. As a starting point of study the structure and behavior of markets, correlation analysis is found to be useful not only in theory but also in practices of portfolio risk estimation and optimization

[? ?]. Especially during the periods of crises, highly collected co-movements of the stocks are very likely to cause significant losses for a portfolio, so it's absolutely necessary to watch the correlations for the portfolio. Also, to understand the market structure and the dynamics, it's interesting to investigate the correlations [? ? ? ? ? ?].

Following the definition and notation widely used in literatures, the *Pearson correlation coefficient* [?]

$$\rho_{ij} = \frac{\langle Y_i Y_j \rangle - \langle Y_i \rangle \langle Y_j \rangle}{\sqrt{\left(\langle Y_i^2 \rangle - \langle Y_i \rangle^2\right)\left(\langle Y_j^2 \rangle - \langle Y_j \rangle^2\right)}} \tag{13}$$

can be calculated for each stock pair of s_i and s_j using the logarithmic return

$$Y_i = \ln P_i(t) - \ln P_i(t-1). \tag{14}$$

The value of ρ_{ij} ranges from -1 to -1 indicating a dynamic relationship for the two stocks from a completely anti-correlation to a completely correlation. For a perfect uncorrelated pair, $\rho_{ij} = 0$ by definition. If there are N stocks in consideration, then there will be N^2 correlation coefficients fitting into a $N \times N$ correlation matrix. Correlation analysis has been applied in the study of market structures [? ? ?] and portfolio optimization [? ? ? ?].

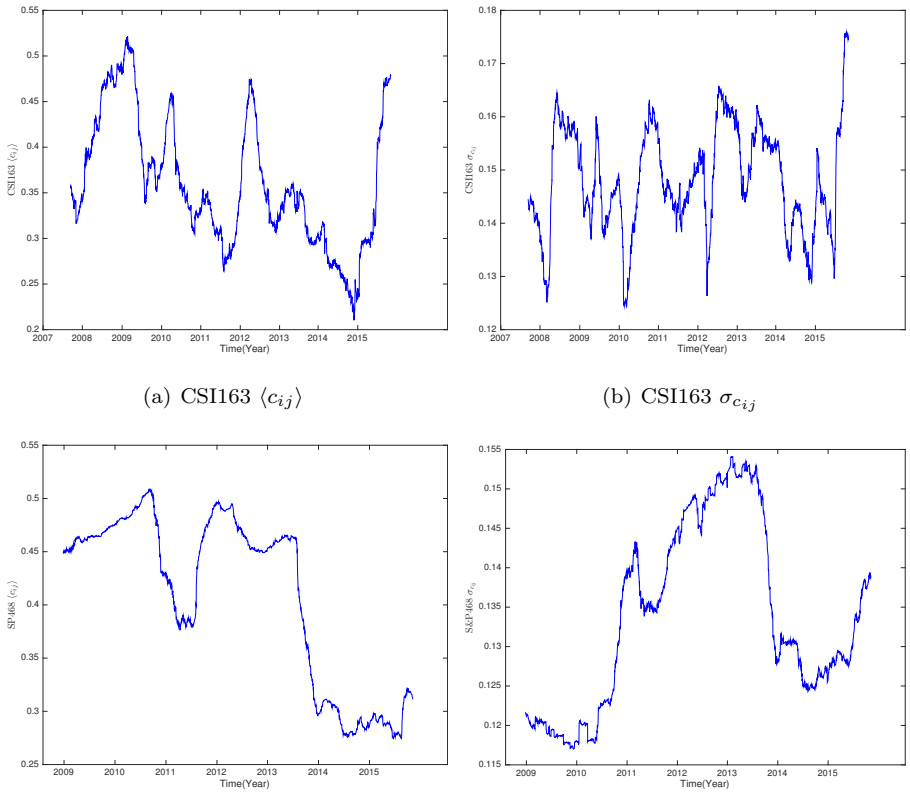
Correlation analysis is so important to be considered as the the basis for many advanced studies such as principal component analysis, singular value decomposition, and factor analysis, etc. But one should not omit the fact that the market is full of noises and the useful information in correlation matrices might be covered by the noises and make correlation analysis less meaningful [?]. To quantify the validations of correlations, recently, there are many work applying *Random Matrix Theory* (RMT) into the studies of the correlation matrices of financial markets [? ? ? ? ? ? ? ? ?]. Rooted from the correlation analysis, RMT offers a new look into the structures and behaviors of the financial markets which are typical complex systems.

In the RMT approach, the statistics of the eigenvalues distribution and the deviation between empirical distribution and the distribution generated from a

random fashion are discussed to describe the information contribution of these deviated eigenvalues and corresponding components of the eigenvectors. But first, the empirical results are tested against a random matrix case [**?**].

4. Correlation properties of CSI163 and S&P468

For each correlation coefficient matrix C, we calculate the average of correlation coefficients $\langle c_{ij} \rangle$ and standard deviation $\sigma_{c_{ij}}$ of both CSI163 and S&P468 and plot the series in Fig. 1(a)-Fig. 1(d).

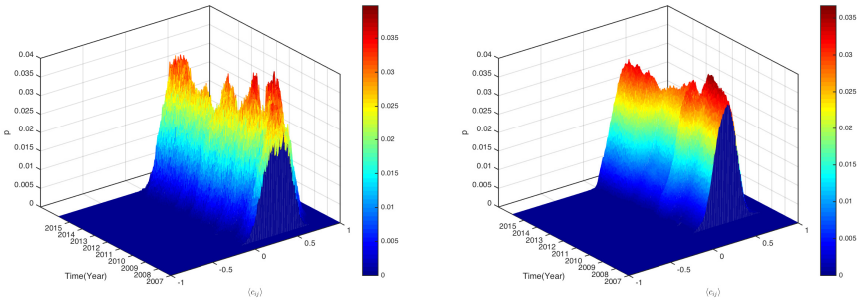


(a) CSI163 $\langle c_{ij} \rangle$ (b) CSI163 $\sigma_{c_{ij}}$

(c) S&P468 $\langle c_{ij} \rangle$ (d) S&P468 $\sigma_{c_{ij}}$

Figure 1: Average of correlation coefficients $\langle c_{ij} \rangle$ and standard deviation $\sigma_{c_{ij}}$ of CSI163 and S&P468 are plotted in Fig. 1(a)-1(d). Since the sliding window size is 500 for S&P468, so the available correlation data starts in the late of 2008.

Then, we investigate the correlation coefficient distributions of the two markets. In Fig. 2(a) and Fig. 2(b), the probability density functions (PDF) of the correlation coefficients $P(c_{ij})$ are plotted for CSI163 and S&P468 respectively. As we can see, the distributions of both markets are obviously shifted away from 0 in positive direction indicating that the markets behavior in a correlated way during most of time and this agrees with the results of [**?**]. To show how the market evolves with the time, in Table 1, we present the overall average of correlation coefficient $\langle c_{ij} \rangle$ and overall standard deviation $\sigma_{c_{ij}}$ for both of CSI163 and S&P468, we see that CSI163 demonstrates a slightly smaller correlation compared to S&P468 which means CSI163 has a larger volatility than S&P468.



(a) CSI163 $\langle c_{ij} \rangle$ (b) S&P468 $\langle c_{ij} \rangle$

Figure 2: Probability density function(PDF) of the correlation coefficients distributions $P(c_{ij})$ of CSI163 (a) and S&P468 (b) evolving with time in our study period.

5. Eigenvalues and eigenvectors for CSI163 and S&P468

5.1. Eigenvalues

Based on the correlation matrices we built in the previous section, we are ready to investigate the eigenvalues and eigenvectors of both markets. First, we use all of the logged daily returns data of both two markets, *ie* CSI163 and S&P468 over the whole study period which is 04/01/2007 and 06/11/2015 covering 2149 trading dates for the former and 2228 trading dates for the latter. We

Table 1: In a sliding window approach, we get the correlation coefficients matrices and we calculate the averages and standard deviations. In this table, we present the overall average, min, and max of $\langle c_{ij} \rangle$ and $\sigma_{c_{ij}}$ for both of CSI163 and S&P468. We can see that, compared to CSI163, S&P468 has a larger value of c_{ij} and smaller values for $\sigma_{c_{ij}}$. This indicates that the stocks of S&P468 are more likely to behavior collectively than CSI163, *ie*, less volatility.

Dataset	$\langle\langle c_{ij} \rangle\rangle$	$\max(c_{ij})$	$\min(c_{ij})$	$\langle \sigma_{c_{ij}} \rangle$	$\max(\sigma_{c_{ij}})$	$\min(\sigma_{c_{ij}})$
CSI163	0.3599	0.5215	0.2103	0.1482	0.1758	0.1243
S&P468	0.4086	0.5084	0.2744	0.1338	0.1541	0.117

present the probability density distributions (PDF) of eigenvalues from empirical correlation matrix and theoretical predicted by using random matrix theory for CSI163 in Fig. 3 and for S&P468 in Fig. 4, respectively. For both markets, we find that most empirical eigenvalues are within the RMT predicted interval with some exceptions. As shown in Fig. 3, for CSI163, the theoretical predicted eigenvalues bounds are $\lambda_{min} = 0.5250$ and $\lambda_{max} = 1.6267$. We see that there are 7 eigenvalues are larger than the largest eigenvalue predicted by RMT, *ie* 4.29% of all eigenvalues fall beyond the interval. The largest eigenvalue $\lambda_1 = 60.2252$ is nearly 37 times of the predicted largest eigenvalue, *ie*, $\lambda_1/\lambda_{\max} = 37.0238$. For S&P468, as shown in Fig. 4, the largest eigenvalue $\lambda_1 = 189.5698$ which is almost 89 times to the bound predicted by RMT. There are 12 eigenvalues are larger than the bound, *ie*, 3.56% are beyond the interval and carry real market information.

Using the sliding window approach, we can investigate the dynamic properties of eigenvalue distributions. For CSI163 and S&P468, we use the window size $L_{csi163} = 170$ and $L_{S\&P468} = 500$ respectively to satisfy the requirement of $Q = L/N > 1$. The study period is between 04/01/2007 and 06/11/2015 covering 2149 trading dates for CSI163 and 2228 trading dates for S&P468. Thus, we have x sliding windows for CSI163 and x for S&P468. As shown in Fig. 5(a) and Fig. 5(b), we see that for both markets, the values of λ_1/N and the average correlation $\langle C_{ij} \rangle$ correlated very well over the whole study period

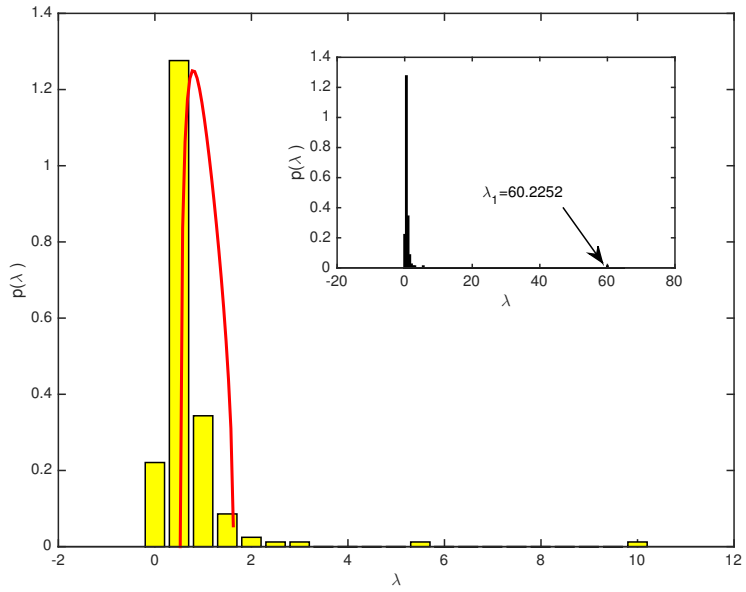


Figure 3: The eigenvalue distributions for CSI163 correlation matrix over the whole study period. The yellow bars are distribution of all eigenvalues calculated from the empirical correlation matrix of 163 daily log return time series and the red curve is the theoretical distribution predicted from the random matrix theory by using a random matrix in the same size of the empirical correlation matrix. The inset is a plot of all empirical eigenvalues including the largest eigenvalue $\lambda_1 = 60.2252$.

indicating that λ_1/N is a good estimator of the average correlation $\langle C_{ij} \rangle$ as we have introduced previously.

In Fig. 6(a) and Fig. 6(b), we plot the largest eigenvalue λ_1/N and the index close prices of CSI300 (a) and S&P500 (b). After the left shifting, we find that λ_1/N and the index itself show similar trends. This shows that λ_1/N is also an indicator of the index itself. For CSI163, the trend similarity is relatively obvious than that of S&P468. If we do not perform left shifts, we find that λ_1/N is anti co-move with the index showing that during markets crashes, the λ_1/N (also the average correlation $\langle C_{ij} \rangle$) gets larger, *ie*, the stocks of the market are correlated, while during calm periods, the λ_1/N gets small indicating less correlations among stocks.

To see how the eigenvalues distributed in the whole study period. In Fig.

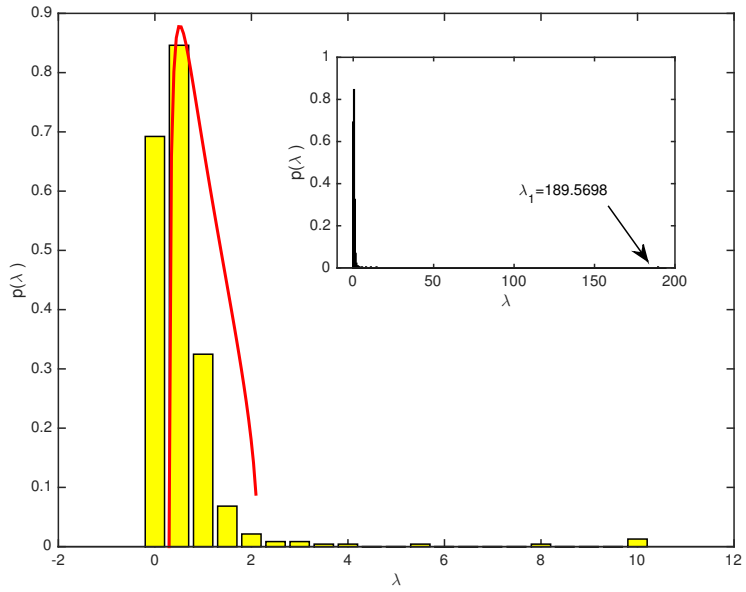
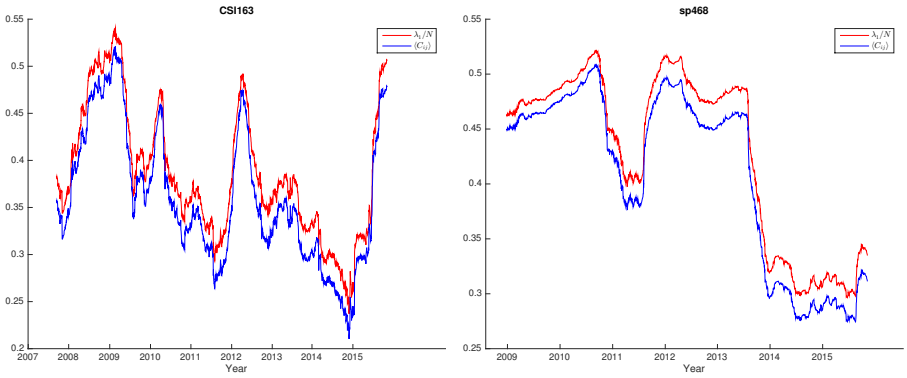


Figure 4: The eigenvalue distributions for S&P468 correlation matrix over the whole study period. The yellow bars are distributions of all eigenvalues calculated from the empirical correlation matrix of 468 daily logged return series and the red curve is the theoretical distribution predicted from the random matrix theory by using a random matrix in the same size of the empirical correlation matrix. The inset is a plot of all empirical eigenvalues including the largest eigenvalue $\lambda_1 = 189.5698$.

7(a)-Fig. 7(b), we plot the distributions of the eigenvalues (excluding the largest eigenvalue) of all sliding windows over the study periods for CSI163 and S&P468, respectively. As the figures show, that most eigenvalues are very small. Though many eigenvalues are within the bounds of prediction based on RMT, we also observe some eigenvalues are larger than the upper bound $\lambda_{max} = 3.9172$ for CSI163 and $\lambda_{max} = 3.8709$ for S&P468. We define the fraction of eigenvalues which are larger than the predicted λ_{max} using RMT as

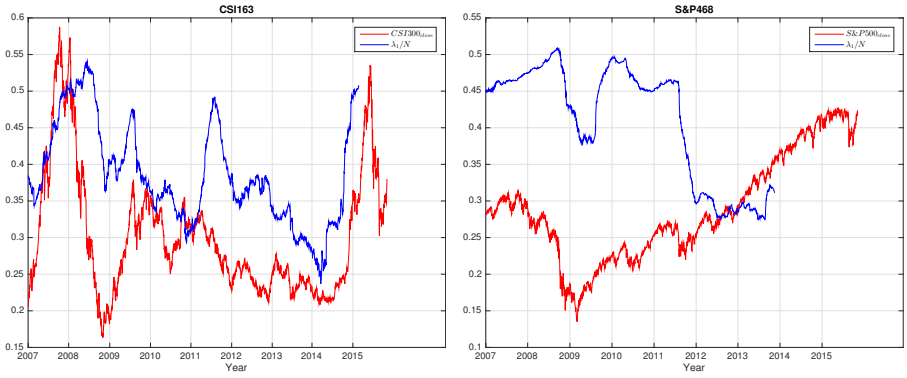
$$p^d = \frac{|\lambda > \lambda_{\max}|}{N}, \tag{15}$$

ie, the ratio of number of eigenvalues deviated beyond λ_{max} to the total number of eigenvalues N. Since the eigenvalues carry meaningful information of the market, this ratio can be treated as an indicator describing how much infor-



(a) CSI163 (b) S&P468

Figure 5: The largest eigenvalue λ_1/N and the average correlation $\langle C_{ij} \rangle$ for all sliding windows of CSI163 (a) and S&P468 (b). We see that the two curves fits very well.

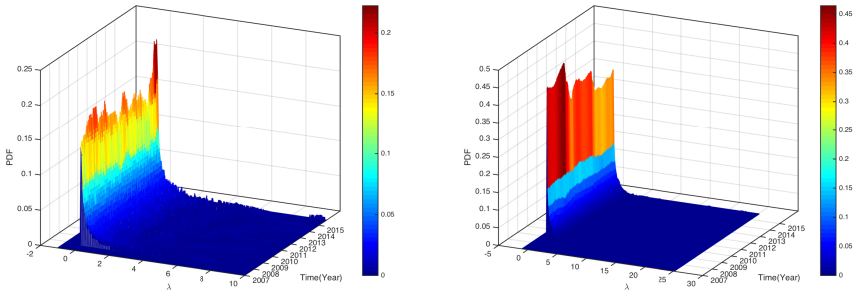


(a) CSI163 (b) S&P468

Figure 6: The largest eigenvalue λ_1/N and the index close price of CSI300(a) and S&P500 (b). The largest eigenvalue λ_1/N curves are left shifted 170 trading dates for CSI300 and 500 trading dates for S&P500, for the window size is 170 for CSI163 and 500 for S&P468. For better visualizations, we shrink the indices of CSI300 and S&P500 with 10000 times and 5000 times respectively. We see that the shifted curves of λ_1/N is similar to the indices.

mation embedded in the empirical eigenvalues distribution. Using the sliding window approach, we calculate the fraction for each window and plot with the index close price for CSI163 in Fig. 8 and S&P468 in Fig. 9, respectively.

For better visualizations, we shrink the index values of 200000 times for



(a) CSI163 (b) S&P468

Figure 7: The PDF of all eigenvalues (excluding the largest eigenvalue λ_1) distributions are plotted for all sliding windows over the study period for CSI163 and S&P468.

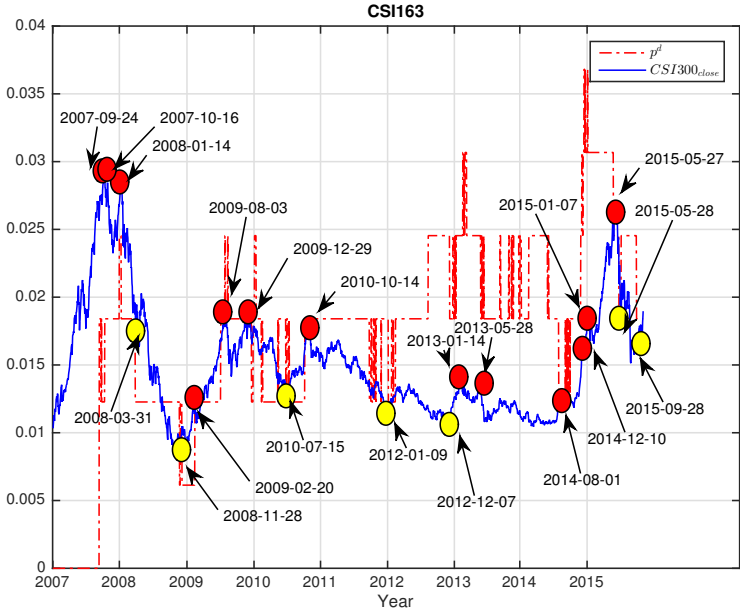


Figure 8: The fraction p^d of eigenvalues beyond the predicted largest eigenvalue versus the index close price for CSI163 over the study period. For better visualization, we rescale the index values by shrinking 200000 times. The coincidences of changes of fraction p^d and the index close price are marked out in red dots for local maximums and yellow dots for local minimums on the price curve with dates.

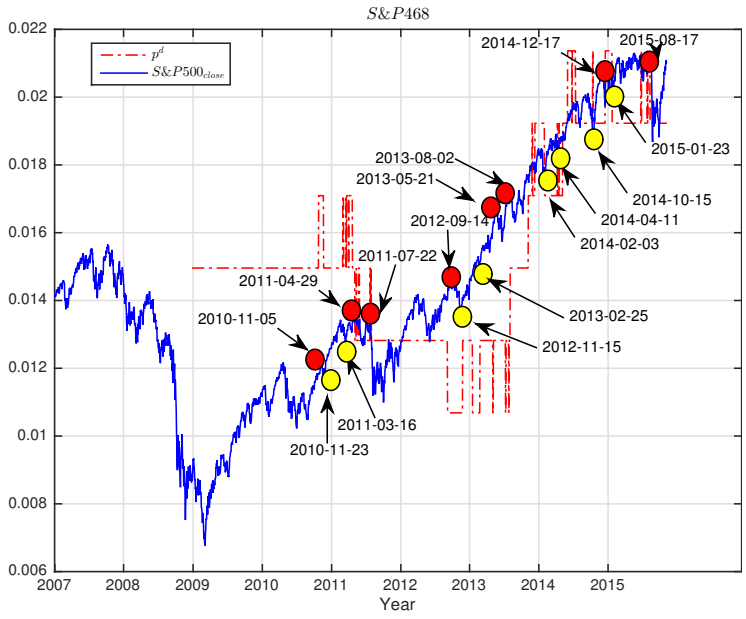


Figure 9: The fraction p^d of eigenvalues beyond the predicted largest eigenvalue versus the index close price for S&P468 over the study period. For better visualization, we rescale the index values by shrinking 100000 times. The coincidences of changes of fraction p^d and the index close price are marked out in red dots for local maximums and yellow dots for local minimums on the price curve with dates.

$CSI163_{close}$ and 100000 times for $S\&P468_{close}$, respectively. As we can see that the values of p^d stay unchanged between sudden changes, so the curves of p^d show a shape of discrete stages with ups and downs. More interestingly, we find that the changes of p^d coincide with the changes of index close prices. As shown in the Fig. 8 for CSI163 and Fig. 9 for S&P468, the changing points of the p^d precisely mark out the local minimums (marked with yellow dots) and local maximums (marked with red dots) of the index itself.

We see that p^d is relatively stable with many unchanged periods but the changes of p^d can match with the major market changes in index close prices, some of them are even leading the index for several days. This indicates that p^d has potentials to monitor the market situation. Once the p^d changed value, investors must be very careful and pay special attention to the market fluctua-

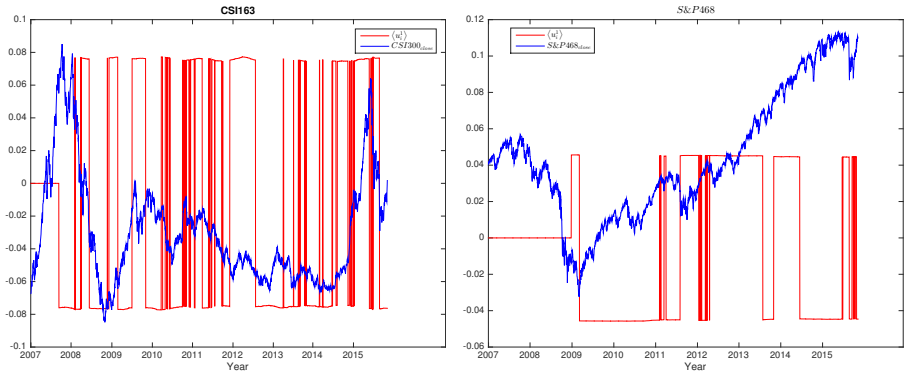
tions both of surges and crashes. This information might be also useful in design trading strategies to catch major market mode switches.

In Table 2, we summarize the properties of eigenvalues which deviate beyond the λ_{max}. We see that only a very small fraction of eigenvalues is larger than the theoretical predicted eigenvalue. On average, only 3.0268 eigenvalues for CSI163 and 7.2250 eigenvalues for S&P468 are beyond the bounds. The average fraction is $\langle p^d \rangle = 0.0186$ for CSI163 and $\langle p^d \rangle = 0.0154$ for S&P468 respectively.

Table 2: Properties of eigenvalue deviation fraction p^d for CSI163 and S&P468. The Avg. Number is the average number of eigenvalues deviated beyond the predicted upper bounds λ_{max}.

Market	Avg. Number	p^d_{min}	p^d_{max}	$\langle p^d \rangle$
CSI163	3.0268	0.0061	0.0368	0.0186
S&P468	7.2250	0.0107	0.0214	0.0154

5.2. Largest eigenvalue



(a) CSI163 (b) S&P468

Figure 10: The average of eigenvector components corresponding to the largest eigenvalue $\langle u_i^1 \rangle$ and the index close price of both CSI300 (a) and S&P500 (b). For better visualizations, we shrink the index close price by 25000 times and 10000 times for CSI300 and S&P500 respectively. We see that the changes of $\langle u_i^1 \rangle$ happen on the dates when the markets changes.

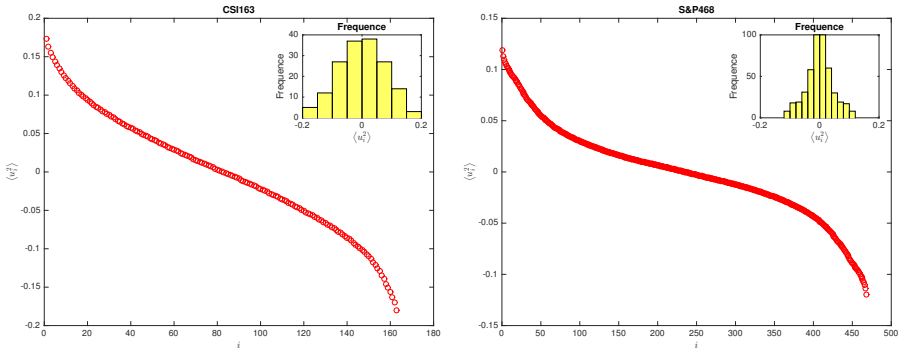
To study the eigenvector u^1 corresponding to the largest eigenvalue λ_1, we take average of all components in the eigenvector. Since the λ_1 stands for the whole market, we expect that the average components are related with the index. We plot the $\langle u_i^1 \rangle$ with the index close prices of both markets for each sliding window in Fig. 10(a) and Fig. 10(b). As shown in the figures, the value of $\langle u_i^1 \rangle$ changes happened on the dates or periods of major market changes. For the eigenvector u^1, we also confirm that all components have the same sign, either positive or negative [**?**], *ie*, all stocks contribute to the movement of the market on the eigenvector u^1 in a same direction, either climb or fall.

5.3. Second largest eigenvalue

It's believed that the largest eigenvalue λ_1 stands for the market mode itself, while the second largest λ_2 eigenvalue and its corresponding eigenvector u^2 contain more information about the market. Now, we focus on the distribution of the components in u^2. As we know that the values of components in eigenvectors represent the weights for the corresponding eigenvector, the best idea to allocate investment in the portfolio management is that we *long* the assets with positive signs and *short* the assets with negative signs, the *eigenportfolio* based on eigenvector u^j is given as:

$$P_j = \sum_{i=1}^{N} \frac{1}{\sqrt{\lambda_j}} \frac{u_i^j}{\sigma_i} Y_i, \tag{16}$$

where N is the number of assets, U_i^j is the component for asset s_i in eigenvector u_j, and Y_i is the return for asset s_i. This indicates that larger eigenvalues λ_i brings less weights for assets in a risky portfolio, while smaller eigenvalues bring smaller risks with greater weights on the asserts. To investigate the components distributions of eigenvectors, we calculate the $u^2(t)$ in a sliding window approach for time t. We first sort eigenvector components in descending order for each sliding window. In Fig. 11(a) and Fig. 11(b), we present the averages for components $\langle u_i^2 \rangle$ of u^2 for CSI163 and S&P468, respectively. We see that both curves share a similar shape and S&P468 has a smaller deviation as shown in the insets.



(a) CSI163 (b) S&P468

Figure 11: The averaged component values $\left\langle u_i^2 \right\rangle$ in eigenvector u^2 for the second largest eigenvalue λ_2 of all sliding windows of CSI163 (a) and S&P468 (b).

To investigate the contribution of stocks for each component, we calculate the numbers of stocks for the corresponding components in u^2. After sorted the components values, we present the results in Fig. 12(a) and Fig. 12(b) for CSI163 and S&P468 respectively. We see that only a few stocks contribute to the components with highest values and smallest values, while for most components, almost all stocks are found presented. The curves are also symmetric. For CSI163, there are 46 stocks appear in the largest component and 45 for the smallest component. For S&P468, there are 24 unique stocks appear in the largest component and 25 for the smallest component. We see that even S&P468 has 468 stocks which is much larger than CSI163, S&P468 has less stocks on the tails on for the largest and smallest components indicating that S&P468 is more concentrated where small group stocks dominate the largest and smallest components.

For industry I_i, the contribution of I_i is defined as

$$I_i^j(t) = \sum_{k \in I_i} \left(u_k^j\right)^2, \tag{17}$$

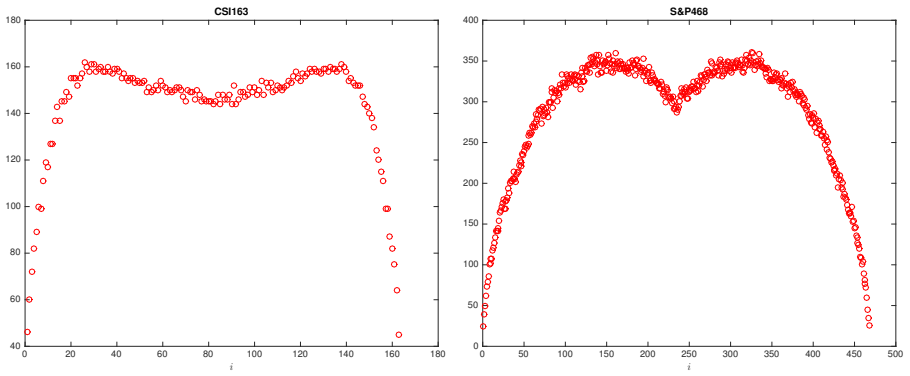
where u_k^j is the value of the stock belonging to industry I_i. By dividing over the

total values of all industries, we get the normalized contribution for industry I_i

$$\overline{I}_i^j(t) = \frac{I_i^j(t)}{\sum_i I_i^j(t)} \tag{18}$$

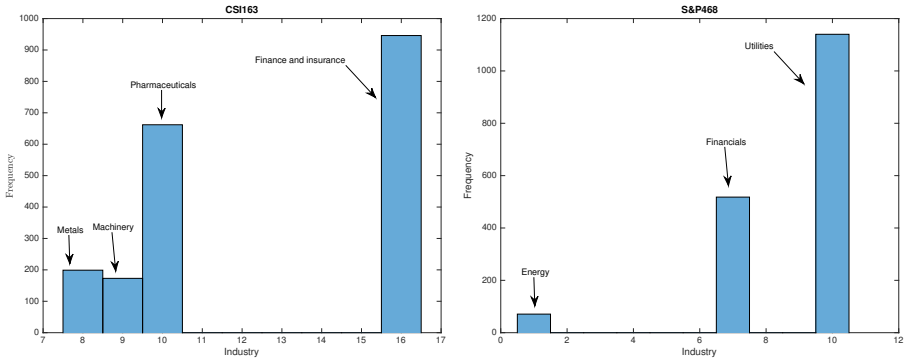
Compared with other approach [**?**], the normalized values allow comparison between any two industries, thus make the ranking of industries possible. Of course, Eq. 18 also indicates that $\sum_i \overline{I}_i^j(t) = 1$.

Using Eq. 17 and Eq. 18, we calculate and rank all industries in all sliding windows for both CSI163 and S&P468. For a given date, we can get the contributions from all sectors to the eigenvector components for the second largest eigenvalue u^2. We investigate which industries appear in the components with largest values. In Fig. 13(a) and Fig. 13(b), we plot the histograms for industries appeared for CSI163 and S&P468, respectively. We find that four industries appeared for CSI163 which are Finance and insurance, Pharmaceuticals, Machinery and Metals, while for S&P468, we find only three industries appeared which are Utilities, Financials, and Energy. This reveals the leading industrial sectors for the two markets over the whole study period.



(a) CSI163 (b) S&P468

Figure 12: The number of unique stocks for all components $\left\langle u_i^2 \right\rangle$ in eigenvector u^2 for the second largest eigenvalue λ_2 of all sliding windows for CSI163 (a) and S&P468 (b). The components are sorted in a descending order.



(a) CSI163 (b) S&P468

Figure 13: The frequencies of industries appearing in the largest values of eigenvector components of CSI163 (a) and S&P468 (b). For CSI163, four industries appear in the largest eigenvector components, while there are three industries for that of S&P468.

5.4. Market switching

Both of the Chinese and US markets experience great fluctuations during our study period covering some major market mode changes of bull markets and bear markets. In our study period between 04/01/2007 and 06/11/2015, the Chinese stock market enjoyed a bull market period from 2007 to 2008 and surged to the historical height in 2008 but soon suffered a major crash and only partially recovered in the middle of 2008 and stepped into bear market mode before long. This bear market mode lasted for almost 7 years only finished in 2015 being replaced by a rocket bull market mode. Unfortunately, the 2015 bull market is very short and tumbled greatly into bear market mode again with huge drops. For S&P500, the US stock markets also suffered a great market crisis in 2008 but the market changed into a very long continuing climbing bull market since 2009.

To investigate how the u^2 changes before and after a market crash, we choose a case study period between 24/07/2008 and 16/02/2009 for CSI300 centering with a market turning point on 04/11/2008 covering 135 trading days and a period between 26/12/2008 and 02/06/2009 for S&P500 centering with a market turning point on 09/03/2009 covering 108 trading days. We denote the ranking

for stock s_i at time t as $R_i(t)$ according to the normalized values. For a period of $[t_s, t_e]$ of length $L_{s,e}$, the averaged ranking for s_i is

$$\langle R_i(t) \rangle = \frac{1}{L_{s,e}} \sum_{t=t_s}^{t_e} R_i(t), \tag{19}$$

where $L_{s,e} = t_e - t_s + 1$ is the number of trading dates in the period. By calculating all averaged rankings for all of the stocks in both periods before and after the market crash, we can get the top and bottom 10 stocks for CSI163 and S&P468. The top and bottom 10 stocks according to the averaged ranking for CSI163 in the *Fall* stage and *Climb* stage are presented in Table 3 and Table 4, respectively. The same lists are presented in Table 5 and Table 6 for the Fall and Climb stages of S&P468.

The tables reveal some very interesting results. In Table 3, we see that stocks of Finance and Real estate occupy the bottom 10 while stocks of Pharmaceuticals dominate the top 10 in the Fall stage of CSI300, and this phenomena remain unchanged during the Climb stage after the market turning point. This indicates and confirms again that Financials are not the only dominate players in Chinese stock market. In the Climbing stage, as shown in the Table 3, stocks of Pharmaceuticals still dominate the top 10 and the stocks of Finance remain at the bottom part. This shows that the internal structure of CSI300 market remain basically unchanged before and after the market crashes.

Being opposite to CSI163, S&P468 demonstrates a different behavior before and after the crash period. As shown in the Table 5, stocks of Financials dominate the top positions with smallest rankings, in other words, stocks of Financials play great roles in the Fall stage, however stocks of Energy collectively occupy the bottom 10. When the market entered the Climb stage passing the turning point, the whole rankings reversed with stocks of Energy turned to be the top stocks while the stocks of Financials fell into the bottom as shown in Table 6.

In Fig. 14(a), we plot the close prices of CSI300 index, Hualan Biological Engineering Inc (Tick 2007, Pharmaceuticals), and Shanghai Pudong Development Bank Co Ltd (Tick: 600000, Finance) in the Fall and Climb stages. In

Table 3: The top ten and bottom ten stocks of the second largest eigenvalue u^2 of CSI163 ranked by the average u^2 components values in the Fall stage between 24/07/2008 and 04/11/2008.

Top 10			
Rank	Tick	Stock Name	Industry
1	2007	Hualan Biological Engineering Inc	Pharmaceuticals
2	600867	Star Lake Bioscience Co Inc	Pharmaceuticals
3	600085	Beijing Tongrentang Co Ltd	Pharmaceuticals
4	963	Huadong Medicine Co Ltd	Wholesale
5	600332	Sichuan Hongda Co Ltd	Metals
6	600108	Gansu Yasheng Industrial (Group) Co Ltd	Agriculture
7	600535	Nanjing Chixia Development Co Ltd	Real estate
8	600277	Jiangsu Hengrui Medicine Co Ltd	Pharmaceuticals
9	600089	TBEA Co Ltd	Machinery
10	999	Sanjiu Medical & Pharmaceutical Co Ltd	Pharmaceuticals
Bottom 10			
Rank	Tick	Stock Name	Industry
154	46	Oceanwide Construction Group Co Ltd	Real estate
155	601988	China Construction Bank	Finance
156	2	China Vanke Co Ltd	Real estate
157	600048	Poly Real Estate Group Co Ltd	Real estate
158	601398	Guangshen Railway	Transportation
159	600016	China Minsheng Banking Corp Ltd	Finance
160	600015	Hua Xia Bank Co Ltd	Finance
161	1	Shenzhen Development Bank Co Ltd	Finance
162	600036	China Merchants Bank Co Ltd	Finance
163	600000	Shanghai Pudong Development Bank	Finance

Table 4: The top ten and bottom ten stocks of the second largest eigenvalue u^2 of CSI163 ranked by the average u^2 components values in the Climb stage between 04/11/2008 and 16/02/2009.

Top 10			
Rank	Tick	Stock Name	Industry
1	999	Sanjiu Medical & Pharmaceutical Co Ltd	Pharmaceuticals
2	2007	Hualan Biological Engineering INC	Pharmaceuticals
3	629	Panzhihua New Steel & Vanadium Co Ltd	Metals
4	600089	TBEA Co Ltd	Machinery
5	600085	Beijing Tongrentang Co Ltd	Pharmaceuticals
6	538	Yunnan Baiyao Industry Co Ltd	Pharmaceuticals
7	963	Huadong Medicine Co Ltd	Wholesale
8	729	Beijing Yanjing Brewery Co Ltd	Food & Beverage
9	600535	Nanjing Chixia Development Co Ltd	Real estate
10	600332	Sichuan Hongda Co Ltd	Metals
Bottom 10			
Rank	Tick	Stock Name	Industry
459	157	Changsha Zoomlion Heavy Industry	Machinery
460	600030	CITIC Securities Co Ltd	Finance
461	600585	Jiangsu Changjiang Electronics Technology	Electronics
462	601988	China Construction Bank	Finance
463	601398	Guangshen Railway	Transportation
464	1	Shenzhen Development Bank Co Ltd	Finance
465	600015	Hua Xia Bank Co Ltd	Finance
466	600016	China Minsheng Banking Corp Ltd	Finance
467	600036	China Merchants Bank Co Ltd	Finance
468	600000	Shanghai Pudong Development Bank Co Ltd	Finance

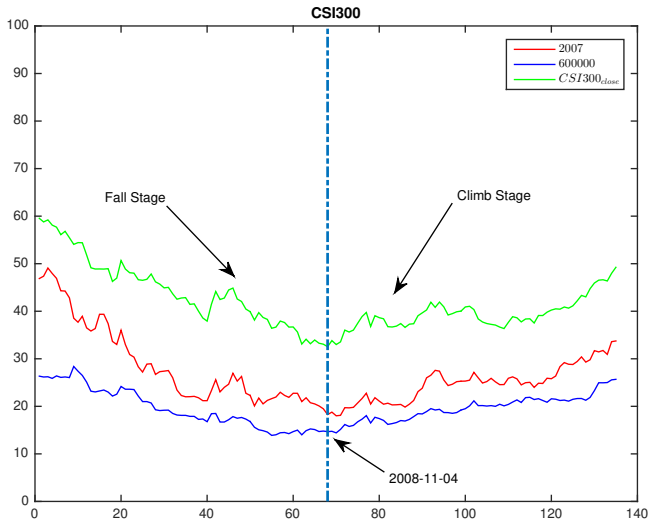
Table 5: The top ten and bottom ten stocks of the second largest eigenvalue u^2 of S&P468 ranked by the average u^2 components values in the Fall stage between 26/12/2008 and 09/03/2009.

Top 10			
Rank	Tick	Stock Name	Industry
1	STI	SunTrust Banks	Financials
2	ZION	Zions Bancorp	Financials
3	MTB	M&T Bank Corp.	Financials
4	CMA	Comerica Inc.	Financials
5	WFC	Wells Fargo	Financials
6	BBT	BB&T Corporation	Financials
7	JPM	JPMorgan Chase & Co.	Financials
8	RF	Regions Financial Corp.	Financials
9	LEN	Lennar Corp.	Consumer Discretionary
10	PNC	PNC Financial Services	Financials
Bottom 10			
Rank	Tick	Stock Name	Industry
459	EOG	EOG Resources	Energy
460	MUR	Murphy Oil	Energy
461	OXY	Occidental Petroleum	Energy
462	HP	Helmerich & Payne	Energy
463	NBL	Noble Energy Inc	Energy
464	XEC	Cimarex Energy	Energy
465	APC	Anadarko Petroleum Corp	Energy
466	DO	Diamond Offshore Drilling	Energy
467	DVN	Devon Energy Corp.	Energy
468	APA	Apache Corporation	Energy

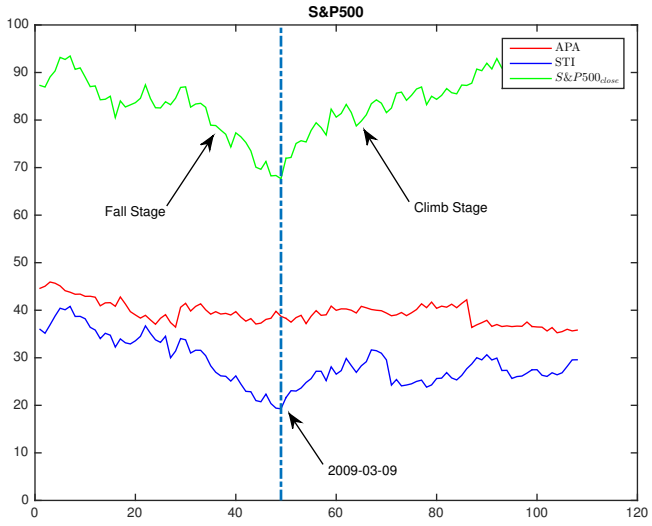
Table 6: The top ten and bottom ten stocks of the second largest eigenvalue u^2 of S&P468 ranked by the average u^2 components values in the Climb stage between 10/03/2009 and 02/06/2009.

Top 10			
Rank	Tick	Stock Name	Industry
1	APA	Apache Corporation	Energy
2	DVN	Devon Energy Corp.	Energy
3	ETR	Entergy Corp.	Utilities
4	DO	Diamond Offshore Drilling	Energy
5	NBL	Noble Energy Inc	Energy
6	APC	Anadarko Petroleum Corp	Energy
7	FE	FirstEnergy Corp	Utilities
8	OXY	Occidental Petroleum	Energy
9	MUR	Murphy Oil	Energy
10	XOM	Exxon Mobil Corp.	Energy
Bottom 10			
Rank	Tick	Stock Name	Industry
459	USB	US Bancorp	Financials
460	JPM	JPMorgan Chase & Co.	Financials
461	RF	Regions Financial Corp.	Financials
462	WFC	Wells Fargo	Financials
463	BBT	BB&T Corporation	Financials
464	PNC	PNC Financial Services	Financials
465	ZION	Zions Bancorp	Financials
466	CMA	Comerica Inc.	Financials
467	MTB	M&T Bank Corp.	Financials
468	STI	SunTrust Banks	Financials

Fig. 14(b), we plot the close prices of S&P500 index, Apache Corporation (Tick: APA, Energy), and SunTrust Banks (Tick: STI, Financials). These four stocks are the top and bottom stocks as listed in the Table 3-Table 6.



(a) CSI163



(b) S&P468

Figure 14: Before and after a market crash of CSI163 (a) and S&P468 (b). For better visualization, the close price of CSI300 is rescaled by 50 times and S&P500 is rescaled by 10 times.

6. Conclusions and discussions

In this study, we applied random matrix theory to study the eigenvalues and their eigenvectors of US and Chinese stock markets. The correlation properties are studied and some eigenvalues of the correlation matrices beyond the predicted bounds are observed in both markets. The largest eigenvalues λ_1 are dozens times larger than the predicted $\lambda_{\max}$. They are found to be as potential market indicators. Eigenvalue deviation fraction beyond the predicted largest eigenvalue are observed to pinpoint market turning points. For the two markets, the most influential industry sectors are identified. They behave differently when market crashes. These findings provide information of the dynamics of eigenvalues and eigenvectors. This is useful for investors and regulators to monitor the markets. On the other hand, the eigenvalues have deep linkages with factor models. The largest eigenvalue stands for the market itself and the corresponding eigenvector has impacts of most stocks, described as the single factor model for stock s_i: $r_i = \beta_i r_M + e_i$, where r_M is the market return, for N stocks, the correlation matrix has one dominant eigenvalue. The CAMP is a special case of single factor model. However, other eigenvalues are beyond the predicted $\lambda_{\max}$. It's natural to model the returns in multi-factors as proposed in Arbitrage Pricing Theory (APT), $r_i = \sum_k \beta_{ki} f_k + e_i$, where f_k is the kth factor. Since the eigenvalues embedded in the predicted bounds represent noises, it's natural to choose the top k largest eigenvalues $\lambda_{\max -k}, \cdots, \lambda_{\max -1}$, thus we get k corresponding eigenvectors $v_{\max -k}, \cdots, v_{\max -1}$, in other words, the k principle components in the PCA. To simplify the model, it's reasonable to consider the sector information which revealed in the eigenvectors, in other words, the corresponding eigenvector components belonging the the sector are reserved. In further researches, it will be interesting to explore possible trading strategies based on the dynamics of eigenvalues and eigenvectors.

Acknowledgements

The authors gratefully acknowledge the funding from National Natural Science Foundation of China under Grant No. 71572028, Ministry of Science and Technology of China under Grant No. 2017IM010700.

References