Extracting Market Trends from the Cross Correlation between Stock Time Series

Mieko Tanaka-Yamawaki
Department of Information and Knowledge Engineering
Graduate School of Engineering, Tottori University,
Tottori, 680-8552 Japan, Japan
mieko@ike.tottori-u.ac.jp

Takemasa Kido
Department of Information and Knowledge Engineering
Graduate School of Engineering, Tottori University,
Tottori, 680-8552 Japan, Japan
s042026@ike.tottori-u.ac.jp

Abstract—**We apply the RMT-PCA, recently developed PCA in order to grasp temporal trends in a stock market, on daily-close stock prices of American Stocks in NYSE for 16 years from 1994 to 2009 and show the effectiveness and consistency of this method by analyzing the whole data of 16 years at once, as well as analyzing the cut data in various lengths between 2-8 years. The extracted trends are consistent to the actual history of the markets. We also discuss on the problem of setting the effective border between the noise and signals considering the artificial correlation created in the process of taking log-return in analyzing the price time series.**

Keywords - RMT-PCA; Correlation; Eigenvalues; Principal Component; Stock Market; Trend.

I. INTRODUCTION

Recently, there have been wide interests on the use of RMT (Random Matrix Theory) in many fields of sciences [1-10]. In particular, the use of asymptotic formula of the eigenvalue spectrum of cross correlation matrix between independent time series of random numbers [11,12], as a reference to the corresponding spectrum derived from a set of different stock price times series in order to extract principal components effectively in a simple way [13-16], has attracted much attention in the community of econo-physics [17, 18]. The main advantage of this method as a principal component analysis is its simplicity. While the standard PCA (Principal Component Analysis) tells us to find the largest PC (Principal Component) and subtract this component from the entire data, and apply the same procedure recursively on the remaining data one by one, RMT-PCA (RMT-based principal component analysis) can process all the "non-random" components at once by subtracting the RMT formula from the eigenvalue spectrum of cross correlation matrix. Plerau, et al. [13] was one of the first attempts to apply this technique on stock price time series. By using the daily close stock prices of NYSE/S&P500, they successfully extracted eminent stocks out of massive data of price time series.

However, this method suffers from two difficulties. One is the restriction on the dimensionality, N, and the length of the data, T, such that N < T. Moreover, the entire set of N times T data are needed for analysis, since the basic quantity of analysis is the cross correlation matrix whose elements are the equal-time inner-products between a pair of stocks. Another difficulty is the restriction of the parameter size. Since the RMT formula is derived in the limit of N and T being infinity, we need a special care to keep the range of the parameters in which the RMT formula is valid.

By using machine-generated random numbers, such as rand(), etc., we have tested the validity of the RMT formula in various range of N and T, and have clarified that N=300, or larger, is the safe range unless T is not too close to N, and the validity decreases for smaller N, and the borderline is around 50<N<100. Since the size of stocks dealt in the major markets exceeds 400, the applicability of RMT formula is justified.

Due to the restriction of the methodology to prepare the length of the time series, T, larger than the dimension of the correlation matrix, N, all the data extending to several years had to be combined into a single correlation matrix in [3-6], in which daily-close prices were used. Thus it was difficult to pin-point a short term trend or to compare trends of different time periods.

By employing intra-day (tick-wise) data containing all the transactions made every day, we can apply the methodology to the data of every year and compare the results of different years. We carried out the same line of study used in [13,14] by setting up the algorithm of RMT-PCA to be applied on intra-day equal-time price correlations. Based on this approach, we have shown that this handy methodology works well to extract the trend change of 4 year interval, from 1994 to 2002 [19,9].

In this paper, we apply the same algorithm to a wider set of stock price data including daily-close prices of American stocks in the database of S&P500 for 16 years from 1994 to 2009. We prepare the data of various lengths by cutting the 16 years into 2, 4 and 8 pieces and check the consistency and effectiveness of the proposed methodology.

II. EIGENVALUE PROBLEM OF CORRELATION MATRIX FOR STOCK PRICES

We shall briefly review the outline of the methodology used in RMT-PCA. The first step is to prepare the price time series into an N×(T+1) matrix named S, whose i-th row contains the price time series of length T+1. This matrix S is converted into a matrix of log-return as follows.

$$r(t) = \log(S(t + \Delta t)) - \log(S(t)) \qquad (1)$$

We normalize each time series in order to have the zero mean the unit variances as follows.

$$x_i(t) = \frac{r_i(t) - < r_i >}{\sigma_i} \quad (i=1,\ldots,N) \qquad (2)$$

The correlation $C_{i,j}$ between two stocks, i and j, can be written as the inner product of the two log-profit time series, $x_i(t)$ and $x_j(t)$,

$$C_{i,j} = \frac{1}{T}\sum_{t=1}^{T} x_i(t)x_j(t) \qquad (3)$$

Here, the suffix i indicates the time series on the i-th member of the total N stocks.

The correlations defined in Eq. (3) makes a symmetric ($C_{i,j} = C_{j,i}$), square matrix whose diagonal elements are all equal to one ($C_{i,i} = 1$) and off-diagonal elements are in general smaller than one ($|C_{i,j}| \leq 1$). As is well known, a real symmetric matrix C can be diagonalized by a similarity transformation $V^{-1}CV$ by an orthogonal matrix V satisfying $V^t=V^{-1}$, each column of which consists of the eigenvectors of C. Such that

$$C v_k = \lambda_k v_k \quad (k=1,\ldots,N) \qquad (4)$$

where the coefficient λ_k is the k-th eigen-value and v_k is the k-th eigenvector. This can be pursued by means of well-known Jacobi's rotation algorithm.

A criterion proposed in [3-6] and examined recently in many real stock data is to compare the result to the formula derived in the random matrix theory [1]. According to RMT, the eigenvalue distribution spectrum of matrix C made of random time series is given by the following formula [2],

$$P_{RMT}(\lambda) = \frac{Q}{2\pi}\frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\lambda} \qquad (5)$$

in the limit of $N \to \infty,\ T \to \infty, Q = T/N = \mathrm{const}$ where T is the length of the time series and N is the total number of independent time series (i.e. the number of stocks considered). This means that the eigenvalues of correlation matrix C between N normalized time series of length T distribute in the following range.

$$\lambda_- < \lambda < \lambda_+ \qquad (6)$$

Following the formula Eq. (5), between the upper bound and the lower bound given by the following formula.

$$\lambda_\pm = (1 \pm Q^{-1/2})^2 \qquad (7)$$

The proposed criterion in our RMT_PCM is to use the components whose eigenvalues, or the variance, are larger than the upper bound λ_+ given by RMT.

$$\lambda_+ < \lambda \qquad (8)$$

III. Application of RMT-PCA on Stock Prices

We prepare N normalized stock returns of the same length T, which makes a rectangular matrix of $S_{i,k}$ where i=1,…,N represents the stock symbol and k=1,…,T represents the traded time of the stocks. The i-th row of this price matrix corresponds to the price time series of the i-th stock symbol, and the k-th column corresponds to the prices of N stocks at the time k. We summarize the algorithm that we used for extracting significant principal components.

However, a detailed analysis of the eigenvector components tells us that the random components do not necessarily reside below the upper limit of RMT, λ_+, but percolate beyond the RMT due to extra randomness added in the process of computing the log-return in Eq. (1). Based on extensive numerical analysis, this percolation always occurs and the maximum front of the continuum spectrum extends to about 20% larger than the upper limit λ_+ of RMT. This fact suggests us that the upper limit λ_+ is not appropriate to separate the signal from the noise due to the percolation of the random spectrum over λ_+ but an effective upper bound λ_{eff} =1.2 λ_+ about 20% larger than the upper limit λ_+ of RMT. Then λ_+ in the step (4) of the RMT-PCA algorithm in Fig. 2 is to be replaced by λ_{eff}.

Algorithm of RMT-PCM :

(1) Select N stock symbols for which the traded price exist for all t=1,…,T, corresponding to all the working days of that term.

(2) Compute log-return r(t) for the selected N stocks. Normalize the time series to have mean=0, variance=0, for each stock symbol, i=1,…, N.

(3) Compute the cross correlation matrix C and obtain eigenvalues and eigenvectors.

(4) Select eigenvalues larger than λ_+, the upper limit of the RMT spectrum, and $\lambda_\pm = (1 \pm Q^{\ 1/2})^2$,

$$P_{RMT}(\lambda) = \frac{Q}{2\pi\lambda}\sqrt{(\lambda_+ - \lambda)(\lambda_- - \lambda\ )}$$

and identify those eigenstates as the principal components.

(5) Sort the eigenvector components corresponding to the eigenvalues identified in the step (4) above, in the descending order and identify the business sectors of the largest 20 components. If those 20 components belong to any particular sector, that is the leading sector in that term.

Figure 1. The algorithm to extract the significant principal components in RMT-PCA.

IV. Trends Extracted as the Eminent Components of Eigenvectors

We applied the algorithm stated in Section 3 on the daily-close prices of American stocks listed in S&P500, for 16 years from 1994 to 2009.

At first, the entire data of this period are used for analysis. Then the entire data is split to 2 parts, 1994-2001 and 2002-2009. Those are further split to 4 parts, 1994-1997, 1998-2001, 2003-2005, 2006-2009. Finally, they are split to 8 parts of 2years data, 1994-1995, 1996-1997, 1998-1999, 2000-2001, 2002-2003, 2004-2005, 2006-2007, 2008-2009. The results are listed in Table 1.

TABLE I. RESULTS FOR 16, 8, 4 YEAR DATA (EIGENVALUES LARGER THAN $2\lambda_+$ ARE HIGHLIGHTED IN BOLD=ITALIC)

	94-09	**94-01**	**02-09**	**94-97**	**98-01**	**02-05**	**06-09**
N	373	373	464	373	419	464	468
T	3961	2015	1946	1010	1002	1006	936
Q	10.6	5.40	4.19	2.71	2.17	2.17	2
λ_+	1.7	2.1	2.2	2.6	2.8	2.8	2.9
λ_1	***74***	***41***	***150***	***37.2***	***53***	***116***	***200***
λ_2	***11***	***13***	***15***	***8.7***	***19***	***14***	***18***
λ_3	***8.8***	***8.8***	***12***	***5.8***	***13***	***13***	***14***
λ_4	***7.7***	***6.9***	***11***	4.6	***9.2***	***9.1***	***8.9***
λ_5	***5.1***	***4.8***	***6.5***	3.3	***6.6***	***6.3***	***5.3***
λ_6	***4.3***	***4.2***	***5.1***	3.2	***5.8***	5.3	5.0
λ_7	3.3	3.5	3.8	2.8	4.7	4.8	4.4
λ_8	2.9	3.1	3.4	2.6	4.2	4.6	3.5
λ_9	2.5	2.7	3.3	2.4	3.8	4.0	3.2
λ_{10}	2.4	2.2	2.8	2.4	3	3.3	2.7
λ_{11}	2.0	2.2	2.4	2.3	2.8	2.9	2.7
λ_{12}	1.9	2.1	2.3	2.3	2.7	2.9	2.5

According to the step (4) in the RMT-PCA algorithm in Fig. 1 and , those 14 eigenstates are the principal components, based on . We find the business sectors of the companies of 20 largest components in the corresponding eigenvectors. If those components are concentrated in any particular business sector, we identify that sector as the trend makers during that time period. It can be proved mathematically that the eigenvector of the largest eigenvalue is consist of components of the same sign, and the corresponding sectors are not concentrated to a particular sector but distributed to any sectors, because the largest principal component show the global feature of the market thus corresponds to its representative index, such as S&P500, in our case of dealing with American stocks. The eigenvectors of the other eigenvalues have components of both signs. It has been known that the positive components and the negative components belong to the two separate business sectors, if they are strongly concentrated to particular sectors. Summing up those knowledge we have, the 2nd principal component reflects the trend of the time period of the data if any concentration of the sectors are observed.

The sectors are classified according to GICS (Global Industry Classification Standard) coding system, that classifies the business sectors of stocks into 10 categories. We denote them by a single capital letter, A-J as follows.

A: Energy, B: Materials, C: Industrials, D: Service,
E: Consumer Products, F: Health Care, G: Financials,
H: Information Technology, I: Telecommunication,
and J: Utility.

If we take λ_{eff} instead of λ_+, as we explained in the last paragraph of Section 3, then we have 10 eigenstates corresponding to the eigenvalues λ_1=74.3,..., λ_{10} = 2.41, we have lesser number of principal components than the above-stated 14. However, the concentration of business sectors in the eigenvector components occurs only for the 4-5 largest eigenvalues and quickly becomes blur for smaller eigenvalues. Based on this observation, we might increase λ_{eff} to the range of $\lambda_{\text{eff}} = 2\lambda_+$, 100% larger than the theoretical criterion. In any case, the difference is irrelevant as long as we take only several principal components.We show 8 bars corresponding to

$v_2(+)$, $v_2(-)$, $v_3(+)$, $v_3(-)$, $v_4(+)$, $v_4(-)$, $v_5(+)$, $v_5(-)$,

where $v_k(+)/v_k(-)$ indicates the positive-sign part/negative-sign part of the vector of k-th principal component, by partitions corresponding to 10 sectors of A-J, and the corresponding eigenvalues and the sign of the components below each bar.

We observe from the graphs in Fig. 4 that the sector H (InfoTech) dominates the (+) components of v_2 and the sector J (Utility) dominates the (-) components of v_2 .

The result of 8 years data, 1994-2001 and 2002-2009 are shown in Fig. 5, the left figure of which shows the dominance of J (Utility) and H (InfoTech) during the term 1994-2001, and the right figure shows the dominance of A (Energy) and G (Financials) during the term 2002-2009. This means the active sector has changed from J (Utility) and H (InfoTech) to A (Energy) and G (Financials) at the turn of the century.

The results of 4 year data, 1994-1997, 1998-2001, 2002-2005, and 2006-2009 are in Fig.6, showing the dominance of J (Utility) and H (InfoTech) both in 1994-1997 and 1998-2001, the dominance of A (Energy) and H (InfoTech) in 2002-2005, and A (Energy) and G (Financials) dominance in 2006-2009. The corresponding result of 2 year data is shown in Fig. 7. No clear structure is seen after 2002, except weak dominance of G (Financials) and A (Energy).

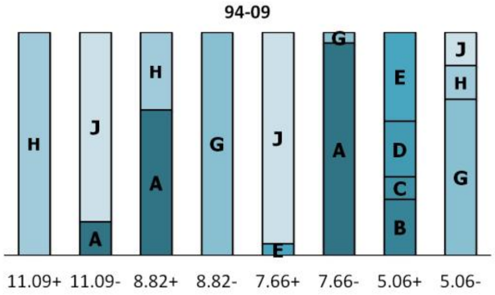
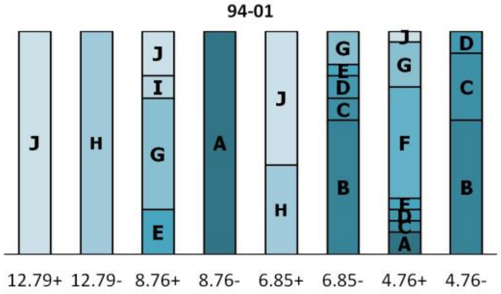


Figure 2. Trends of 16 years from 1994 to 2009 are shown. The sector H (Information Technology) and J (Utility) are the most eminent sectors in this period.



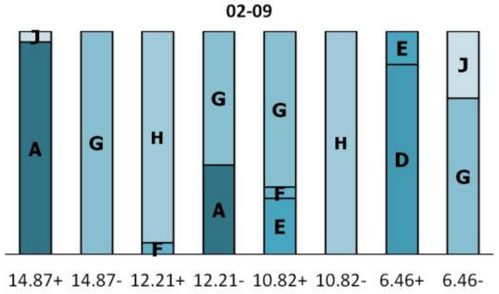
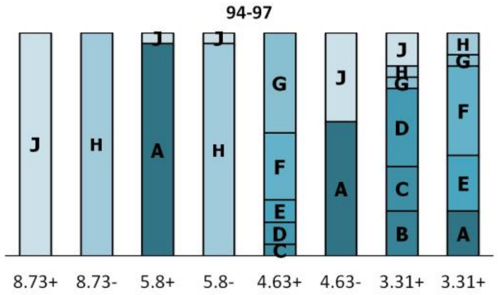
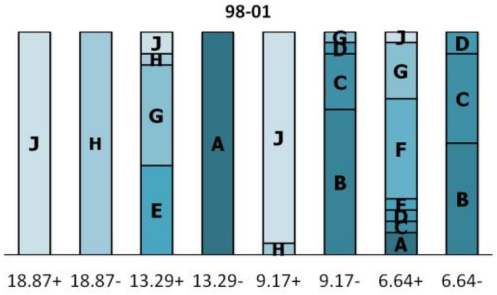
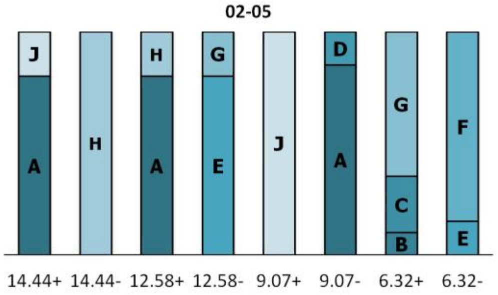


Figure 3. Trends of 8 years, 1994-2001 (left) and 2002-2009 (right). In 1994-2001, the sector J (Utility) and H (Information Technology) dominate, but in 2002-2009, A (Energy) and G (Financial) dominate the market.







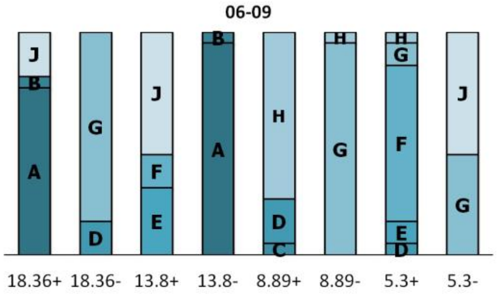
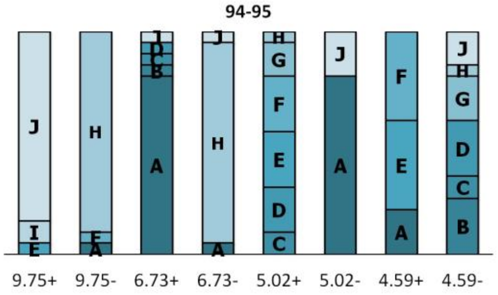
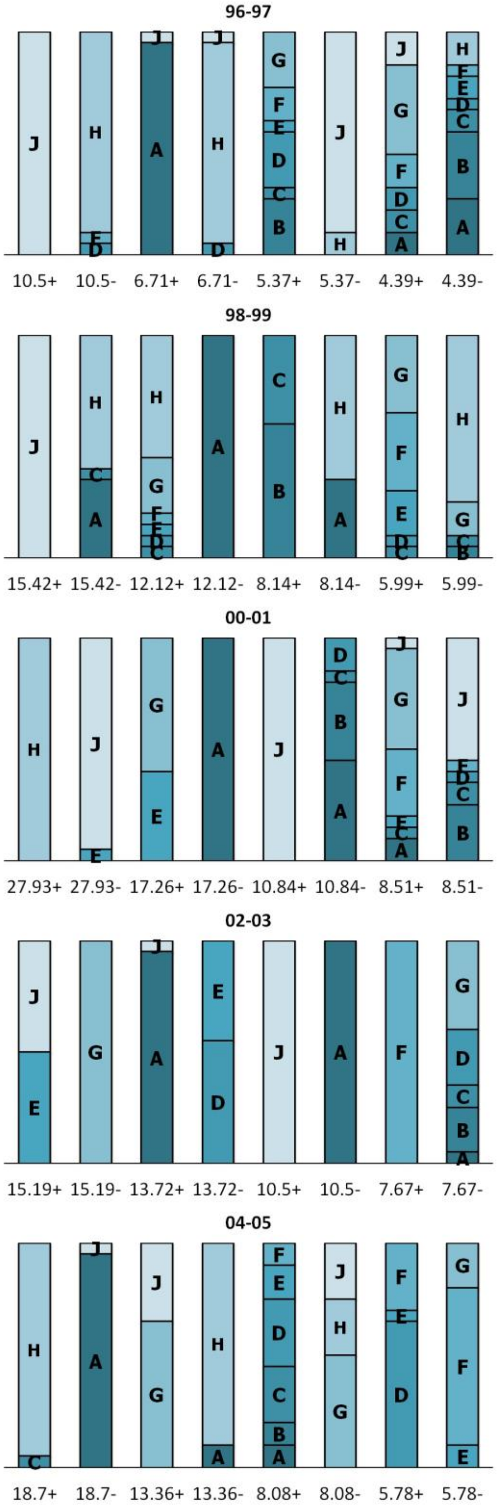


Figure 4. Trends of 4 years each are shown. Both in 1994-1997 and 1998-2001, J (Utility) and H (Information Technology) dominate, while A (Energy) and H (Information Technology)dominate in 2002-2005 and A (Energy) and G (Financial) dominate in 2006-2009.



ISBN: 978-1-61208-154-0



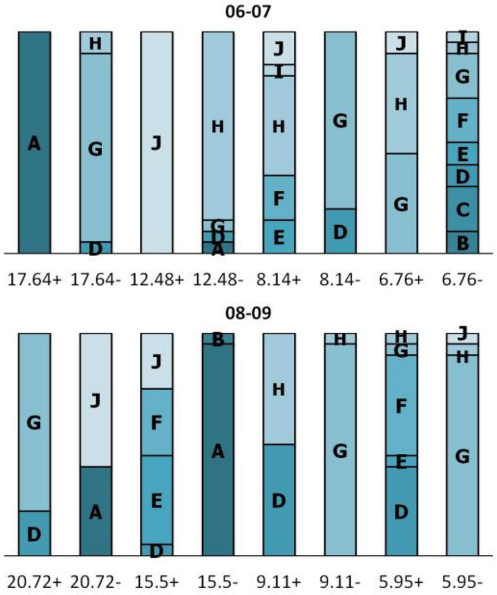


Figure 5. Trends of 4 years each are shown. Both in 1994-1997 and 1998-2001, J (Utility) and H (IT) dominate, while A (Energy) and H (IT) dominate in 2002-2005 and A (Energy) and G (Financial) dominate in 2006-2009.

V. CONCLUTION AND DISCUSSION

Our results have shown that the trend of each time period can be successfully depicted by the concentrated business sectors in the positive components and the negative components of the eigenvector corresponding to the 2nd principal components. Although the condition $\lambda > \lambda_+$ dramatically reduces the number of principal components compared to the conventional method of PCA. Moreover, our method is considerably simple with much shorter in process to extract principal components, which is a great advantage in the case of analyzing the stock market.

The conventional PCA tells us to extract the largest principal component and subtract this element from the entire data, and apply the same procedure recursively on the remaining data one by one. This kind of method requires a lot of computational time and is not suitable for analyzing a system of the large dimension, such as a set of stocks in the market. Another method of PCA uses the eigenvalues of the correlation matrix of times series, but tells us to pick up the components whose eigenvalues are larger than one, or the accumulated sum of eigenvalues exceeds 80 percent of the total sum, etc. Neither one is suitable for analyzing the stocks in the market, since the number of principal components thus obtained usually exceeds 100 for N=400-500, while the RMT- PCA has derived the number of principal components in the range of 5-13 in our lesson in Section 4 in this paper. We illustrate this point in Fig. 8. We also tabulate the numbers of principal components obtained in our analysis applied on 2 years data in Table 1.

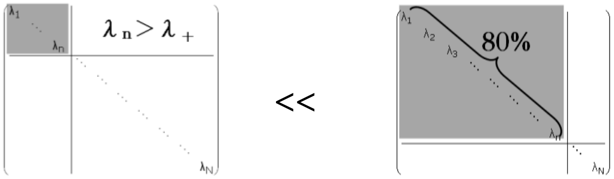


Figure 6. The RMT-PCA (left) offers much smaller number of PCs compared to the method of 80 percent accumulative eigenvalues (right).

Some remarks are in order before concluding this paper. Firstly, we mention the limitation of finantial data. The daily-close price data can be dowloaded on the web site, such as Yahoo Finance, etc. with free of charge, thus convenient for us to analyze. Moreover, the data are ready to use to calculate the equal-time correlation. Althogh some stocks are not traded every day, more than 400 stock symbols are traded every day in NYSE and TOPIX. However, the length of data per year is only 252, the working days of the markets. Thus we must combine two or more years to satisfy $N < L$.

On the other hand, the intra-day price data are sold commertially and quite costly. Moreover, the traded time for each stocks are not the same and we need pre-process the data in order to make the equal-time correlation. We have chosen the "block-tick" method to regard the trades within a certain block, such as every hour, to make them equal-time.

Finally, we mention the applicability of RMT-PCA on other fields of study. This methodology is not restricted for studying stock prices but can be applied for much wider variety of noisy data, including meteorological, or atomospheric data, and demographical data. Since the equal-time cross correlation matrix is the main tool of the methodology, we need to prepare a large number of equal-time data taken simultaneously at every moments, whose length L is larger than the number of the time series N, namely, $N < L$. Moreover, $N > 300$ is desirable by the reason explained in Section I.

Considering the time-delay for any information to propagates, however, the equal-time correlation is not sufficient, and we eventually need to consider the correlation with time delay. Technically speaking, however, the time-delayed correlation matrix C is not symmetric and the eigenvalues are not guaranteed to be real valued. The eigenvalue problem of such matrices is much more complicated compared to the equal-time correlation matrix, which is symmetric thus can be diagonalized by using the Jacobi's rotation algorithm.

References

[1]. Mehta, M. L., “Random matrices”, 3rd edition, Academic Press (2004)

[2]. Edelman, Alan and Rao, N. Raj, "Random matrix theory", Acta Numerica, pp. 1-65 (2005)

[3]. Bai, Zhidong and Silverstein, Jack, "Spectral Analysis of Large Dimensional Random Matrices", Springer (2010)

[4]. Tao, Terence and Vu, Van, "Random matrices: universality of ESD and the circular law" (with appendix by M. Krishnapur), Annals of Probability, Vol. 38 (5), pp. 2023-2065 (2010)

[5]. Beenakker, C. W. J., "Random-matrix theory of quantum transport", Reviews of Modern Physics, Vol. 69, pp. 731–808 (1997)

[6]. Kendrick, David, "Stochastic control for economic models", McGraw-Hill (1981)

[7]. Bahcall, S. R., "Random matrix model for superconductors in a magnetic field", Physical Review Letters, Vol. 77, pp. 5276–5279 (1976)

[8]. Franchini, F., Kravtsov, V. E., "Horizon in random matrix theory, the Hawking radiation, and flow of cold atoms", Physical Review Letters, Vol.103, 166401 (2009).

[9]. Peyrache, A., Benchenane, K., Khamassi, M., Wiener, S. I., and Battaglia, F. P., "Principal component analysis of ensemble recordings reveals cell assemblies at high temporal resolution", Journal of Computational Neurosience, Vol. 29, pp. 309–325 (2009)

[10]. Sánchez, D., Büttiker, M., "Magnetic-field asymmetry of nonlinear mesoscopic transport". Physical Review Letters, Vol. 93, 106802 (2004).

[11]. Marcenko, V A, Pastur, L A, "Distribution of eigenvalues for some sets of random matrices", Mathematics of the USSR-Sbornik, Vol. 1, pp. 457–483 (1994)

[12]. Sengupta, A. M. and Mitra, P. P., “Distribution of singular values for some random matrices”, Physical Review E, Vol. 60, pp. 3389-3392 (1999)

[13]. Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L.A.N., and Stanley, H.E., “Random matrix approach to cross correlation in financial data”, Physical Review E, Vol. 65, 066126 (2002)

[14]. Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L.A.N., and Stanley, H. E., "Scaling behaviour in the growth of companies", Physical Review Letters, Vol. 83, pp. 1471-1474 (1999)

[15]. Laloux, L., Cizeaux, P., Bouchaud, J.-P., and Potters, M., "Noise dressing of financial correlation matrix", Physical Review Letters, Vol. 83, pp. 1467-1470 (1999)

[16]. Bouchaud, J.-P. and Potters, M., “Theory of Financial Risks”, Cambridge University Press (2000)

[17]. Mantegna, R .N. and Stanley, H. E., “An Introduction to econophysics: correlations and complexity in finance”, Cambridge University Press (2000)

[18]. Iyetomi, H. et al., Fluctuation-dissipation theory of input-output interindustrial relations, Physical Review E, Vol. 83, 016103 (2011)

[19]. Tanaka-Yamawaki, Mieko, "Extracting principal components from pseudo-random data by using random matrix theory", Lecture Notes in Artificial Intelligence, Vol. 6278, pp. 602- 611 (2010)

[20]. Tanaka-Yamawaki, Mieko, "Cross correlation of intra-day stock prices in comparison to random matrix theory", Intelligent Information Management (online journal: http://www.scrp.org) (2011)