ELSEVIER

Physica A 287 (2000) 374–382

PHYSICA A

www.elsevier.com/locate/physa

A random matrix theory approach to financial cross-correlations

V. Plerou*, P. Gopikrishnan, B. Rosenow,
L.A.N. Amaral, H.E. Stanley

Center for Polymer Studies, Department of Physics, Boston University, Boston, MA 02215, USA

Received 13 May 2000; received in revised form 4 July 2000

Abstract

It is common knowledge that any two firms in the economy are correlated. Even firms belonging to different sectors of an industry may be correlated because of "indirect" correlations. How can we analyze and understand these correlations? This article reviews recent results regarding cross-correlations between stocks. Specifically, we use methods of random matrix theory (RMT), which originated from the need to understand the interactions between the constituent elements of complex interacting systems, to analyze the cross-correlation matrix **C** of returns. We analyze 30-min returns of the largest 1000 US stocks for the 2-year period 1994–1995. We find that the statistics of approximately 20 of the largest eigenvalues (2%) show deviations from the predictions of RMT. To test that the rest of the eigenvalues are genuinely random, we test for universal properties such as eigenvalue spacings and eigenvalue correlations, and demonstrate that **C** shares universal properties with the Gaussian orthogonal ensemble of random matrices. The statistics of the eigenvectors of **C** confirm the deviations of the largest few eigenvalues from the RMT prediction. We also find that these deviating eigenvectors are stable in time. In addition, we quantify the number of firms that participate significantly to an eigenvector using the concept of inverse participation ratio, borrowed from localization theory. © 2000 Published by Elsevier Science B.V. All rights reserved.

Keywords: Random matrix theory; Cross-correlations; Econophysics

1. Introduction

Increasing evidence for scale-invariant distributions and correlation functions in financial time series have attracted the attention of several physicists [1,2]. One reason for this interest is the quest for mechanisms that gives rise to scale invariance. In physical systems, scale-free behavior [3,4] is often caused by correlations that become long

* Corresponding author.
E-mail address: plerou@cgl.bu.edu (V. Plerou).

0378-4371/00/$ - see front matter © 2000 Published by Elsevier Science B.V. All rights reserved.
PII: S0378-4371(00)00376-9

range in the vicinity of a 'critical' value of a tunable parameter such as temperature. Can we understand scale invariance in financial data in similar terms? To answer this question, it is important to analyze and quantify the nature of correlations between different units comprising the system. Thus, the problem of understanding correlations between the different units that comprise the market is of scientific interest. Furthermore, a precise quantification of correlations between the returns of different stocks is of practical importance in quantifying the risk of portfolios of stocks [2,5,6], pricing of options, and forecasting.

The problem of quantifying correlations between the price changes of different stocks can be expressed using the following simple problem. Consider a box containing many gas molecules and suppose there is some mechanism which records the velocities of each of the gas molecules. Next, suppose that there are some random *pair-wise* bonds between some of the gas molecules. How can we identify, which molecules are connected? The problem is simply solved: we start by calculating from the records of velocities v_i of molecules $i = 1,\ldots,N$, a cross-correlation matrix $C_{ij} \equiv \langle v_i v_j \rangle - \langle v_i \rangle \langle v_j \rangle / \sigma_i \sigma_j$, where σ_i denotes the standard deviation of v_i and $\langle \cdots \rangle$ denotes the time average from the entire time series of v_i. If we had infinitely many records for v_i, we would identify the non-zero off-diagonal entries C_{ij} which would correspond to the pairs i, j that were connected.

To complicate the problem by one more level, suppose that we do not have just random pair-wise bonds, but rather bonds connecting clusters of molecules. How can we find from the records of the velocities, which clusters are connected? One approach is to identify the principal components or eigenvalues (and eigenvectors) of the matrix C_{ij}. The participants of the eigenvectors of C_{ij} would contain information about the clusters of connected molecules, similar to the problem of a N-body system, interconnected by springs, where the eigenvectors of the Hamiltonian contain information about the different modes of oscillation.

Next, suppose that the clusters that are connected by bonds do not just stay the same in time, but rather evolve, i.e., new molecules are connected to the already-existent clusters and some molecules which are part of one cluster become part of other clusters. What can we do in such a case? We can still analyze the principal components of the cross-correlation matrix, which would contain information about which molecules on the average remained in a particular cluster for the period of time analyzed. If the stability of these bonds in time is low, then we would expect the measured correlations C_{ij} to be mostly random.

Finally, if we add to the problem, finite length of time series used for computing the matrix elements C_{ij}, then it is quite difficult a problem to identify which clusters remained bonded on the average, over the time period analyzed.

2. Correlations between stocks

The problem of identifying stocks that are correlated is not unlike the complex example discussed above. We have time series of price fluctuations G_i for $i = 1,\ldots,N$

stocks [1] from which we calculate the correlation matrix **C** which has elements

$$C_{ij} \equiv \frac{\langle G_i G_j \rangle - \langle G_i \rangle \langle G_j \rangle}{\sigma_i \sigma_j}, \qquad (1)$$

where $\sigma_i \equiv \sqrt{\langle G_i^2 \rangle - \langle G_i \rangle^2}$ is the standard deviation of the price changes of company i, and $\langle \cdots \rangle$ denotes a time average over the period studied. The difficulty in quantifying correlations between any two stocks i, j arises from the following:

- Unlike most physical systems, there is no "algorithm" to calculate the "interaction strength" between two companies i, j (as there is for, say, two spins in a magnet). The problem is that although every pair of company should interact either directly or indirectly, the precise nature of interaction is unknown.
- Correlations need not be just pairwise but rather involving clusters of stocks.
- The correlations C_{ij} between any two pairs i, j of stocks change with time.
- For each stock i, we have only finite records $\{G_{ij}, j = 1, \ldots, T\}$, from which to estimate an average correlation.

3. Why random matrices?

How can we identify the correlated clusters of stocks when there is randomness in the measured correlations C_{ij}, either in the form of correlations that change in time, or by the finite length used to compute the correlation matrix elements? The problem of understanding the properties of matrices with random entries is one which has a rich history originating from 1950 nuclear physics from the work of Wigner, and later on by Dyson and Mehta, and many results are known [7–14]. In the case of nuclear physics, the problem was to understand the energy levels of complex nuclei, when model calculations failed to explain experimental data.

The problem was tackled by Wigner, who made the bold assumption that the interactions between the constituents comprising the nucleus are so complex that they can be modeled as random. As a result, Wigner assumed that the Hamiltonian **H** describing a heavy nucleus has, in the matrix representation, elements H_{ij} which can be assumed as mutually independent random numbers. Based on this assumption alone, Wigner derived [7–9] properties for the statistics of eigenvalues of **H**, which were in remarkable agreement with experimental data.

RMT predictions represent an average over all possible interactions [10–12]. Deviations from the *universal* predictions of RMT identify system-specific, non-random properties of the system under consideration, providing clues about the underlying interactions [13,14].

[1] The data analyzed are the 30-min returns, i.e., $T = 6448$ records of $N = 1000$ stocks from the TAQ data base for the 2-year period 1994–1995.

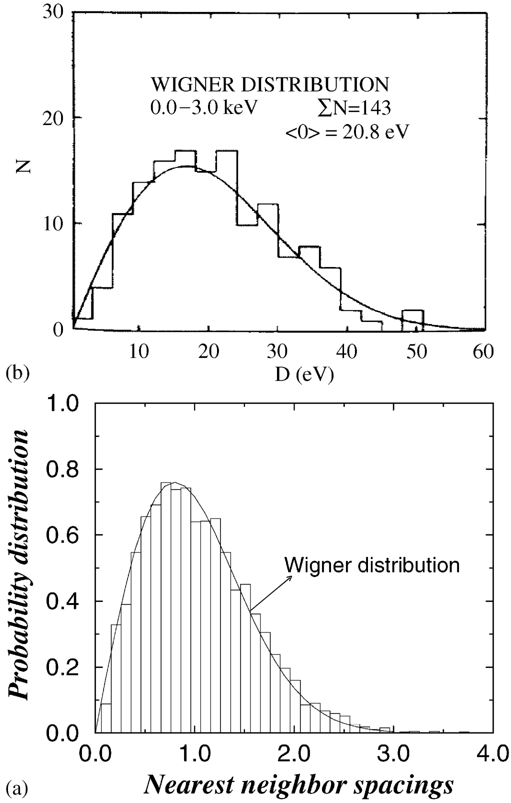

Fig. 1. (a) Histogram of eigenvalue (energy level) spacings, of a heavy nucleus. The solid line shows Wigner's prediction for the energy level spacings of complex nuclei, calculated using the only assumption of a symmetric Hamiltonian matrix with independent random entries [from H.I. Liou et al., Phys. Rev. C 5 (1972) 3; for more examples, see T.A. Brody et al., Rev. Mod. Phys. 53 (1981) 385]. (b) Nearest-neighbor spacing distribution of the eigenvalues of the cross-correlation matrix **C** of stock price fluctuations (from Ref. [15]) after unfolding [13]. The bold line is the Wigner distribution for real symmetric matrices.

The class of matrices Wigner considered are real symmetric matrices, whose elements are distributed according to a Gaussian probability distribution, the Gaussian orthogonal ensemble (GOE). For such matrices, Wigner showed that

$$P_{\mathrm{GOE}}(s) = \frac{\pi s}{2}\exp\left(-\frac{\pi}{4}s^2\right) , \tag{2}$$

where $P_{\mathrm{GOE}}(s)$, some times referred to as Wigner distribution, is distribution of energy level spacings s. This theoretical prediction was later successfully tested on empirical data (Fig. 1a).

Here, we review how this framework can be used to quantify and understand the correlations between different stocks [15,16]. We first compute the eigenvalue spacings $s \equiv \lambda_{i+1} - \lambda_i$, where λ_i denote the rank-ordered eigenvalues after unfolding [13,15], and compare the spacing distribution $P(s)$ to the Wigner distribution $P_{\mathrm{GOE}}(s)$ (Fig. 1b). We find remarkable agreement – thus suggesting that the empirical cross-correlation matrix

C is indeed consistent with a real-symmetric random matrix (or a GOE matrix). We also apply more sensitive tests such as the number variance of λ and find agreement [15] with RMT results for GOE matrices, thus confirming the consistency of **C** with RMT predictions.

What can we infer from this result? From the scientific side, agreement of the eigenvalue statistics of **C** with RMT results implies that **C** has entries that contain a considerable degree of randomness. Randomness could be the result of either nonstationary correlations or a result of the finite time series used. To test that finiteness of time series alone cannot be the reason for RMT agreement, we increase the length of the time series T used to compute **C** by a factor of 4. We still find agreement of the eigenvalue spacing distribution with RMT predictions, suggesting that RMT agreement is also due to nonstationary correlations. From the practical side, RMT agreement of the statistics of **C** argues against the wide use of empirically measured C_{ij} in a variety of applications.

4. Deviations from RMT predictions

The results presented above are universal properties of the correlation matrix that agree well with RMT predictions. Deviations from RMT indicate properties that are specific to the system and arise from the presence of collective modes. For example, deviations of the level spacings of certain nuclei from the Wigner distribution was found to be connected to collective modes of the nucleus [17]. How can we detect collective behavior? One approach is to study the eigenvalue distribution of **C**.

For **C** constructed out of uncorrelated time series, the eigenvalue distribution was calculated exactly [18,19]. We can therefore compare [15,16,20] the distribution $P(\lambda)$ with the prediction for uncorrelated time series. Fig. 2a shows $P(\lambda)$ for **C**. We note that the "bulk" of the distribution is consistent with the RMT bounds calculated in Refs. [18,19]. This comparison also indicates the presence of several eigenvalues clearly outside the random matrix bound (Fig. 2a). Particularly interesting is the largest eigenvalue, which is approximately 25 times larger than the value predicted for a random correlation matrix – suggesting genuine information about the correlations between different stocks.

Having demonstrated that the bulk of the eigenvalues satisfies RMT predictions, we now proceed to analyze the eigenvectors of **C**. We first analyze the statistics of the eigenvectors [15,16]. The distribution of eigenvector components for a random correlation matrix is a Gaussian with zero mean and unit variance. An examination of the eigenvectors corresponding to the eigenvalues which deviate from the random-matrix bound shows systematic deviations from the Gaussian prediction. In particular, the largest eigenvalue is strongly non-Gaussian, tending to uniform (Fig. 2b) – suggesting that all companies participate. This notion can be accurately quantified by the concept of inverse participation ratios, borrowed from the localization theory, where we find indeed that all components participate approximately equally to the largest eigenvector [15]. This implies that every company is connected with every other company. In the

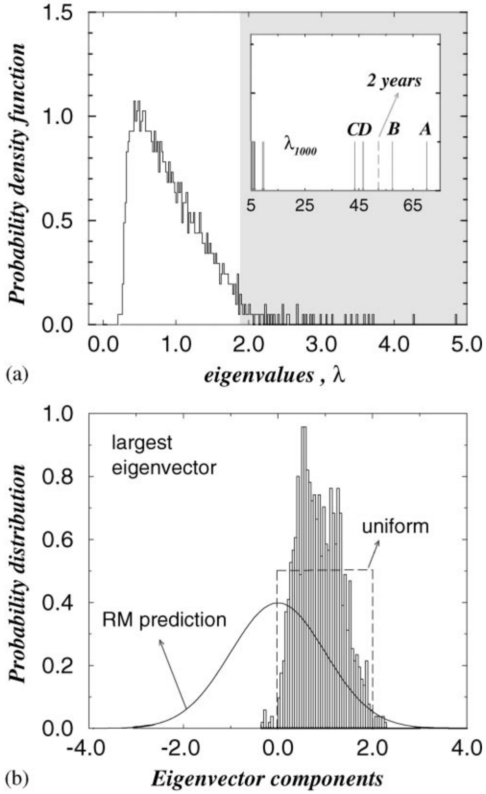

Fig. 2. (a) The probability density of the eigenvalues of the normalized cross-correlation matrix **C** for the 1000 largest stocks in the TAQ database for the 2-year period 1994–1995. Analytical results predict the eigenvalue distribution within a bound $0.37 \leqslant \lambda_k \leqslant 1.94$ for the eigenvalue distribution of a cross-correlation matrix [18,19] from uncorrelated time series. There are several eigenvalues in the shaded region, outside the random matrix bound. The inset shows the largest eigenvalues for correlation matrices computed for four six-month periods in 1994–1995, denoted as A, B, C, and D. The dashed point shows the largest eigenvalue $\lambda_{1000} \approx 50$ for the entire two years, which is approximately 25 times larger than the maximum eigenvalue predicted for uncorrelated time series. (b) The distribution of eigenvector components for the eigenvalues within the RMT bound show agreement with Gaussian behavior whereas the eigenvalues outside the RMT bound show significant deviations from the Gaussian prediction of RMT, which implies "collective" behavior or correlations [13] between different companies. The largest eigenvalue would then correspond to the correlations within the entire market [15,16]. Shown is the distribution of eigenvector components corresponding to the largest eigenvalue, which conforms to an approximately uniform distribution with all companies contributing.

stock market problem, this eigenvector conveys the fact that the whole market "moves" together and indicates the presence of correlations that pervade the entire system.

In addition, we also examine the stability in time of the eigenvectors corresponding to the eigenvalues that deviate from RMT bounds. To test the time stability, we first split the entire two year period into four six-month sub-periods A, B, C, and D. For each sub-period, we calculate a cross-correlation matrix, and compute its eigenvalues and eigenvectors. We then identify, from each sub-period, approximately 15 largest

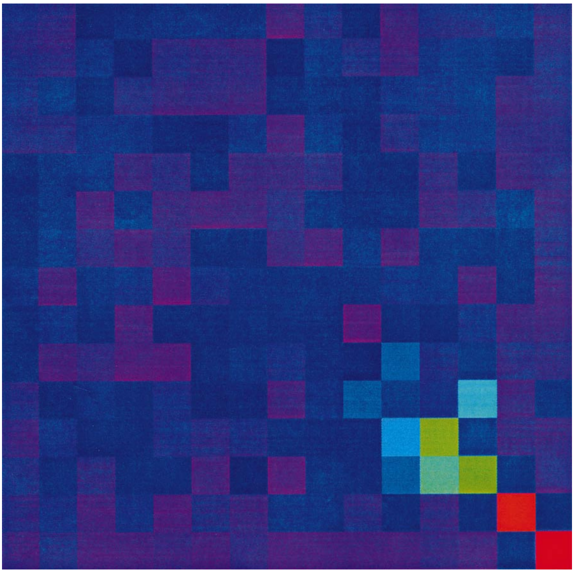

Fig. 3. Color-coded pixel representation of O_{ij} which shows the scalar product of the eigenvectors corresponding to the 15 largest eigenvalues from period A and B (six months apart): here, left to right on the horizontal axis denote the 15 largest eigenvalues of period A in ascending order, and similarly, on the vertical axis, top to bottom show 15 largest eigenvalues of period B in ascending order. The color coding is done such that blue corresponds to 0 and red corresponds to 1. The eigenvectors corresponding to the largest four eigenvalues show considerable stability (even for larger time scales of approximately 1 year). The rest of the eigenvectors toward the RMT bounds (toward the left on the horizontal axis and top on the vertical axis) show less stability.

eigenvectors that deviate from the RMT bounds. Let us denote by a_i, $i = 1,\dots,15$, the 15 eigenvectors of period A (in ascending order of eigenvalue), and similarly b_j, $j = 1,\dots,15$, for period B. We measure time stability by the scalar product

$$O_{ij}(\tau) \equiv \sum_{\ell=1}^{N} a_{i\ell}(t) b_{\ell j} , \tag{3}$$

where O_{ij} is a 15×15 matrix, and $N = 1000$ is the number of components of each eigenvector. If the vectors are perfectly stable, then we expect O_{ij} to be diagonal with elements 1. No stability would mean all elements of O_{ij} are zero. We show in Fig. 3, a color-coded version of O_{ij}, which shows that the eigenvectors corresponding to the largest 4,5 eigenvalues show large values of O_{ij}. As we move toward the RMT bound, the eigenvectors show decreasing amounts of stability.

5. Correlations and scaling?

In the previous sections, we presented evidence for different modes of correlations between different companies. For example, the largest eigenvalue of the cross-correlation matrix showed correlations that pervade the entire market. In physical systems under certain conditions, long-range correlations between subunits result in scale-invariant properties of the system [3,4]. Could it be that the above observed cross-correlations result in scale-invariant behavior?

Recent studies [21] show that the distribution of returns for individual companies and for the S&P 500 index have the same asymptotic power-law behavior with an exponent $1+\alpha \approx 4$. This is surprising because the distribution of index returns $G_{\mathrm{SP500}}(t)$ does not show convergence to Gaussian behavior – even though the 500 distributions of individual returns $G_i(t)$ that form $G_{\mathrm{SP500}}(t)$ are not statistically stable. More precisely,

$$G_{\mathrm{SP500}}(t) \equiv \sum_{i=1}^{500} w_i G_i(t) , \tag{4}$$

where $w_i \equiv S_i / \sum_{j=1}^{N} S_j$, where S_i denotes the market capitalization of company i. From the central limit theorem for random variables with finite variance, we expect that the probability distribution of $G_{\mathrm{SP500}}(t)$ would show signs of convergence to Gaussian, provided there are no significant dependencies among the returns G_i for different i. Instead, it is found that the distribution of $G_{\mathrm{SP500}}(t)$ has the same asymptotic behavior as that for individual companies.

In Ref. [22], it was shown that when $G_i(t)$ is time-shuffled, the observed scaling behavior between S&P 500 index returns and individual stock returns breaks down – suggesting the existence of non-trivial cross-correlations that cause scale-invariant behavior. Using RMT methods, we have seen that the largest eigenvalue is by far the strongest influence common to all stocks. Therefore, in the spirit of the often-used market models, one possible way to reconcile the largest eigenvalue is to express

$$G_i(t) = \alpha_i + \beta_i M(t) + \varepsilon_i(t) , \tag{5}$$

where $M(t)$ represents an influence common to all stocks, α_i, and β_i are parameters that can be estimated by a regression, and $\varepsilon_i(t)$ is a stock-specific term. Using this simple market model, one can attribute the scaling behavior of S&P 500 and individual stock returns to the market $M(t)$, which is a common influence for all stocks.

Recent studies have also analyzed economic data from the physics perspective of a complex system with each unit depending on the other. Stanley and Salinger first located and secured a database – called COMPUSTAT – that lists the annual sales of all publicly traded firms in the United States. With this information, Stanley and co-workers calculated histograms of how firm sizes change from one year to the next [23]. They found that the distribution of growth rates of firm sales has the same functional form regardless of industry or market capitalization. Moreover, the width of these distributions σ decrease with increasing size S (measured by sales) as a power law with an exponent approximately 1/6. Similar scaling exponents were also found for different measures of size S such as the number of employees. Recently, similar statistical properties were found for the GDP of countries [24] and for university research fundings [25].

The scaling behavior $\sigma(S) \sim S^{-\beta}$, with $\beta \approx 1/6$ is surprising, since, one expects by the central limit theorem that $\beta = 1/2$ [23]. Hence, it is not impossible to imagine that the value of $\beta \approx 1/6$ found for such diverse systems may be the result of correlations, similar to those we found for stock prices, that involve all subunits of the system, because similar empirical laws appear to hold for data on a range of systems that at first sight might not seem to be so closely related.

Acknowledgements

We conclude by thanking all our colleagues from whom we learned a great deal, especially, S.V. Buldyrev, D. Canning, P. Cizeau, T. Guhr, S. Havlin, R.N. Mantegna, C.-K. Peng, M.A. Salinger, and M.H.R. Stanley.

References

[1] R.N. Mantegna, H.E. Stanley, An Introduction to Econophysics: Correlations and Complexity in Finance, Cambridge University Press, Cambridge, 1999.
[2] J.P. Bouchaud, M. Potters, Theory of Financial Risk, Cambridge University Press, Cambridge, 2000.
[3] H.E. Stanley, Introduction to Phase Transitions and Critical Phenomena, Oxford University Press, Oxford, 1971.
[4] H.E. Stanley, Scaling, universality, and renormalization: three pillars of modern critical phenomena, Rev. Mod. Phys. 71 (1999) S358.
[5] G.W. Kim, H.M. Markowitz, J. Portfolio Management 16 (1989) 45.
[6] E.J. Elton, M.J. Gruber, Modern Portfolio Theory and Investment Analysis, Wiley, New York, 1995.
[7] E.P. Wigner, Ann. Math. 53 (1951) 36.
[8] E.P. Wigner, Proc. Cambridge Philos. Soc. 47 (1951) 790.
[9] E.P. Wigner, Conference on Neutron Physics by Time-of-Flight, Gatlinburg, Tennessee, 1956.
[10] F.J. Dyson, J. Math. Phys. 3 (1962) 140.
[11] F.J. Dyson, M.L. Mehta, J. Math. Phys. 4 (1963) 701.
[12] M.L. Mehta, F.J. Dyson, J. Math. Phys. 4 (1963) 713.
[13] M.L. Mehta, Random Matrices, Academic Press, Boston, 1991.
[14] T. Guhr, A. Müller-Groeling, H.A. Weidenmüller, Phys. Rep. 299 (1998) 190.
[15] V. Plerou, P. Gopikrishnan, B. Rosenow, L.A.N. Amaral, H.E. Stanley, Phys. Rev. Lett. 83 (1999) 1471.
[16] L. Laloux, P. Cizeau, J.-P. Bouchaud, M. Potters, cond-mat/9810255, Phys. Rev. Lett. 83 (1999) 1469.
[17] S.S. Drozdz, S. Nishizaki, J. Speth, M. Wójcik, Phys. Rev. E 57 (1998) 4016.
[18] F.J. Dyson, Revista Mexicana de Física 20 (1971) 231.
[19] A.M. Sengupta, P.P. Mitra, Phys. Rev. E 60 (1999) 3389.
[20] S. Galluccio, J.-P. Bouchaud, M. Potters, Physica A 259 (1998) 449.
[21] P. Gopikrishnan, V. Plerou, L.A.N. Amaral, M. Meyer, H.E. Stanley, Phys. Rev. E 60 (1999) 5305.
[22] V. Plerou, P. Gopikrishnan, L.A.N. Amaral, M. Meyer, H.E. Stanley, Phys. Rev. E 60 (1999) 6519.
[23] M.H.R. Stanley, L.A.N. Amaral, S.V. Buldyrev, S. Havlin, H. Leschhorn, P. Maass, M.A. Salinger, H.E. Stanley, Nature 379 (1996) 804.
[24] Y. Lee, L.A.N. Amaral, D. Canning, M. Meyer, H.E. Stanley, Phys. Rev. Lett. 81 (1998) 3275.
[25] V. Plerou, L.A.N. Amaral, P. Gopikrishnan, M. Meyer, H.E. Stanley, Nature 400 (1999) 433. (cond-mat/9906229).