Random matrix theory analysis of cross correlations in financial markets

Akihiko Utsugi,[1] Kazusumi Ino,[2] and Masaki Oshikawa[1]
[1]*Department of Physics, Tokyo Institute of Technology, Oh-okayama 2-12-1, Meguro-ku, Tokyo, 152-8551, Japan*
[2]*Department of Pure and Applied Sciences, University of Tokyo, Komaba 3-8-1, Meguro-ku, Tokyo, 153-8902, Japan*
(Received 5 December 2002; revised manuscript received 12 November 2003; published 24 August 2004)

We confirm universal behaviors such as eigenvalue distribution and spacings predicted by random matrix theory (RMT) for the cross correlation matrix of the daily stock prices of Tokyo Stock Exchange from 1993 to 2001, which have been reported for New York Stock Exchange in previous studies. It is shown that the random part of the eigenvalue distribution of the cross correlation matrix is stable even when deterministic correlations are present. Some deviations in the small eigenvalue statistics outside the bounds of the universality class of RMT are not completely explained with the deterministic correlations as proposed in previous studies. We study the effect of randomness on deterministic correlations and find that randomness causes a repulsion between deterministic eigenvalues and the random eigenvalues. This is interpreted as a reminiscent of "level repulsion" in RMT and explains some deviations from the previous studies observed in the market data. We also study correlated groups of issues in these markets and propose a refined method to identify correlated groups based on RMT. Some characteristic differences between properties of Tokyo Stock Exchange and New York Stock Exchange are found.

DOI: 10.1103/PhysRevE.70.026110 PACS number(s): 89.65.−s, 05.40.Fb

I. INTRODUCTION

The price changes of securities such as stocks involve various economic backgrounds as well as interaction between securities. They seem to be quite complicated. Conventionally, financial economists model the price changes of securities by stochastic processes (*random walks*) [1]. It is a basic ingredient of modern portfolio theory [2]. Although the use of stochastic processes is common in finance, the validity of such a formulation should be empirically tested, e.g., by statistical properties of the markets, since the underlying ergodic property of a market may be hard to be established.

Recently, the statistical characterizations of financial markets based on physics concepts and methods attract considerable attention [3]. Given that a stochastic model is valid, some statistical properties of the market should be derived as outsets of stochasticity. For example, the cross correlation matrix among N securities can be regarded as a random matrix and it may be legitimate to expect that it shares universal properties of a corresponding ensemble of random matrix theory (RMT) in an appropriate large N limit (since N is usually large). This has been confirmed by several studies on actual stock markets [4–6]. The bulk of the eigenvalue distribution of the cross correlation matrix of a major index [Standard and Poors 500 (S&P 500)] of the New York Stock Exchange (NYSE) is found to follow the eigenvalue distribution of the Wishart matrix [4], which is a random correlation matrix constructed from mutually uncorrelated time series [7,8]. Also the eigenvalue spacing statistics are found to follow those of the gaussian orthogonal ensemble (GOE) [5].

The aim of this paper is to yield further support on the applicability of RMT to analysis of stock markets. In Sec. II, we give a brief review on the relevant results of RMT. We describe our data sample in Sec. III. In Sec. IV, we test predictions of RMT for the cross correlation matrix for the daily prices of the issues in the Tokyo Stock Exchange (TSE) from 1993 to 2001. The quantities we calculated are the distribution of the eigenvalues, the nearest- and next-nearest-neighbor spacings, rigidity, and a certain moment of eigenvector components. We find good agreement with the real data within the RMT bounds for the eigenvalues. Indeed, there are clear deviations outside the bounds which indicate the presence of deterministic correlations among issues. In Sec. V, we consider random variables with deterministic correlations and show that the bulk part of the eigenvalue distribution of the correlation matrix is stable. In Sec. VI, we closely examine the distribution of the moment of eigenvector component. Eigenvectors corresponding to the eigenvalues outside the RMT bounds deviate from the RMT prediction. According to Ref. [6], the deviating eigenvalues at the lower edge are a consequence of the strong correlations among a few issues. However, we find that the observed data are not explained quantitatively by this reasoning alone. Therefore, we analyze the effect of randomness on deterministic correlations between issues and find an interplay between deterministic correlations and randomness. We argue that it gives a refined explanation on the deviations. In Sec. VII, we identify groups of strongly correlated issues from the information of the nonrandom eigenvectors. The ways of grouping in the TSE and NYSE show some differences.

II. BRIEF REVIEW ON RANDOM MATRIX THEORY

A. Wishart matrix

Let $S_i(t)$ be a price at time t of a stock labeled by i ($i=1,2,\cdots,N$, $t=1,2,\cdots,T$). The change of price at time t can be measured by

$$G_i(t) \equiv \ln S_i(t+1) - \ln S_i(t). \tag{1}$$

Here, we take logarithm of the prices because the fluctuation of stock prices is typically given by the geometric Brownian motion. Since

$$G_i(t) \simeq \frac{S_i(t+1) - S_i(t)}{S_i(t)}, \tag{2}$$

where $G_i(t)$ is approximately the return of the issue i from t to $t+1$. We also define the normalized return $g_i(t)$ as follows:

$$g_i(t) \equiv \frac{G_i(t) - \langle G_i \rangle_T}{\sigma_i}. \tag{3}$$

$\langle \cdots \rangle_T$ indicates the time series average of T steps and the dispersion σ_i is given by

$$\sigma_i \equiv \sqrt{\langle G_i^2 \rangle_T - \langle G_i \rangle_T^2}. \tag{4}$$

Then, the correlation matrix C is expressed in terms of $g_i(t)$

$$C_{ij} \equiv \langle g_i g_j \rangle_T, \tag{5}$$

where C is a real symmetric matrix with positive eigenvalues.

We will model the price of stocks as a stochastic process. For N random variables $x_i(t)$ $(i=1,2,\cdots,N)$, a matrix M which is defined by $M_{ti}=x_i(t)$ is a $T\times N$ matrix. The cross correlation matrix W is defined as follows:

$$W_{ij} \equiv \langle x_i x_j \rangle_T = \frac{1}{T} M^t M, \tag{6}$$

where M^t is the transposition of M. A purely random case with a uniform dispersion σ is given by

$$\langle x_i(t) \rangle = 0, \tag{7}$$

$$\langle x_i(t) x_j(\tau) \rangle = \sigma^2 \delta_{ij} \delta_{t\tau}. \tag{8}$$

Here $\langle \cdots \rangle$ indicates the average over the random variable phase space. In this case, W is called the Wishart matrix [7,8]. We can include "true" correlations among issues by replacing δ_{ij} in Eq. (8) by a nondiagonal matrix $\widetilde{C}$. We will call $\widetilde{C}$ a *deterministic correlation* while we call C or W a *cross correlation*.

B. Eigenvalue statistics of random matrices

Let us summarize the relevant results of RMT to which we will refer in this paper.

In the limit $N\to\infty, T\to\infty$ with $Q\equiv T/N$ fixed, the eigenvalue distribution $\rho(\lambda)$ for the Wishart matrix becomes [9]

$$\rho(\lambda) = \frac{Q}{2\pi\sigma^2} \frac{\sqrt{(\lambda_{\max} - \lambda)(\lambda - \lambda_{\min})}}{\lambda}, \tag{9}$$

$$\lambda_{\min}^{\max} = \sigma^2 \left(1 + \frac{1}{Q} \pm 2\sqrt{\frac{1}{Q}} \right), \tag{10}$$

where Eq. (9) is exact at $N\to\infty, T\to\infty$ with $Q\equiv T/N$ = constant. It is approximately valid at finite N and T when N and T are not small. According to Eqs. (9) and (10), the eigenvalues of the Wishart matrix distribute only in the range $(\lambda_{\min}, \lambda_{\max})$.

Next, we consider the Gaussian ensembles of random matrices. In the Gaussian ensembles, the probability of a matrix H to be in the infinitesimal volume element dH (dH is given by the product of infinitesimal of independent elements) is given by $P(H)dH$ where $P(H)$

$$P(H) = A \exp\left(-a \sum_i |\lambda_i|^2 \right). \tag{11}$$

Here, a is a parameter which characterizes the ensemble, λ_i is the eigenvalue of H, and A is the normalization constant. For general ensembles, one replaces the term $\Sigma_i |\lambda_i|^2$ by $\Sigma_i V(\lambda_i)$ with a function $V(\lambda)$. For example, one can add the quartic or higher-order terms, but it is known that, in the large N-limit (N is the size of H), the model flows to the Gaussian model [10]. The Gaussian models are classified by the symmetry of the matrix as: (i) GOE, the ensemble invariant under the orthogonal group, (ii) Gaussian symplectic ensemble, the ensemble invariant under the symplectic group, and (iii) Gaussian unitary ensemble (GUE), the ensemble invariant under the unitary group. Since the correlation matrix C is real symmetric, the ensemble relevant to our analysis is GOE. For GOE, volume element dH is given by

$$dH = \prod_{i\leqslant j} dH_{ij}. \tag{12}$$

To obtain the statistical measure of the eigenvalue distribution $P(\lambda_1, \lambda_2, \cdots, \lambda_N)$, one expresses H as the product of the diagonal matrix with eigenvalue entries and the other variables, and then integrates the other variables. In this way, we get the measure

$$\prod_{i<j} |\lambda_i - \lambda_j|^\beta \prod_k d\lambda_k. \tag{13}$$

Here, $\beta=1$ for GOE, $\beta=2$ for GUE, and $\beta=4$ for GSE. Thus, the eigenvalue distribution for a Gaussian ensemble is determined by β. By this way, we get the eigenvalue distribution for a general potential V as follows:

$$P(\lambda_1, \lambda_2, \ldots, \lambda_N) = A' \exp\left[-\beta \left(\sum_{k=1}^{N} \frac{V(\lambda_k)}{\beta} - \sum_{i<j} \ln|\lambda_i - \lambda_j| \right) \right], \tag{14}$$

where A' is the normalization constant. From Eq. (14), one sees that the statistical properties at the short spacing between eigenvalues are dominated by $-\ln|\lambda_i-\lambda_j|$ and the total potential is negligible. Thus, β determines the eigenvalue spacing at a short distance. For each β, the level spacing has been closely studied [11]. As the correlation matrix is real symmetric, we expect that its statistical properties of the eigenvalue spacing are given by $\beta=1$. One can characterize the statistical properties of eigenvalue spacing by the nearest-neighbor spacing P_{nn}, the next-nearest-neighbor spacing P_{nnn}, and the "rigidity" $\Delta(L)$. P_{nn} and P_{nnn} are for short-range correlations while $\Delta(L)$ is for long-range correlations. $\Delta(L)$ is defined as

$$\Delta(L) \equiv \frac{1}{L}\left\langle \min_{A,B} \int_{\lambda-\frac{L}{2}}^{\lambda+\frac{L}{2}} (F(\lambda') - A\lambda' - B)^2 d\lambda' \right\rangle_{\lambda}, \quad (15)$$

where $F(\lambda)$ is given by

$$F(\lambda) = \sum_k \Theta(\lambda - \lambda_k), \quad (16)$$

with the Heaviside function Θ. $F(\lambda)$ counts the number of eigenvalues below λ. The meaning of $\Delta(L)$ is that one fits $F(\lambda)$ by a line in an interval with a width L around each eigenvalue, and take the average of the deviations of the fit. $\Delta(L)$ is small when the eigenvalue spacing has a uniform distribution.

For GOE, P_{nn}, P_{nnn}, and $\Delta(L)$ are given by [11],

$$P_{nn}(s) = \frac{\pi s}{2} \exp\left(-\frac{\pi}{4}s^2\right), \quad (17)$$

$$P_{nnn}(s) = \frac{2^{18}}{3^6\pi^3} s^4 \exp\left(-\frac{64}{9\pi}s^2\right), \quad (18)$$

$$\Delta(L) = \frac{1}{15}L^{-4}\int_0^L du(L-u)^3(2L^2 - 9Lu - 3u^2) \times\left(\frac{1}{2}\delta(u) - Y(u)\right), \quad (19)$$

where $Y(u)$ is called a two-spectral cluster function given by

$$Y(u) = \left(\frac{\sin(\pi u)}{\pi u}\right)^2 + \frac{d}{du}\left(\frac{\sin(\pi u)}{\pi u}\right)\int_u^\infty \frac{\sin(\pi t)}{\pi t} dt. \quad (20)$$

According to RMT, the distribution of components of an eigenvector of GOE is the normal distribution with mean 0 and dispersion N. A useful quantity in characterizing the distribution of components is the inverse participation ratio (IPR) [11,12]. For each eigenvector u_k, IPR is defined by the following formula:

$$I_k \equiv \sum_{i=1}^{N} u_{ki}^4, \quad (21)$$

where u_{ki} is the i-th component of u_k. For example, let us consider the case u_{ki} is $1/\sqrt{L}$ for $1 \leqslant i \leqslant L$ and 0 for the other i's. This gives $I_k = 1/L$. Thus IPR can be interpreted as the inverse of the number of components which differ from zero significantly. In RMT, the expectation value of IPR is

$$\langle I_k \rangle = N\int_{-\infty}^{\infty} u_{ki}^4 \frac{1}{\sqrt{2\pi N}} \exp\left(-\frac{u_{ki}^2}{2N}\right) du_{ki} = \frac{3}{N}. \quad (22)$$

III. MARKET DATA

The data we analyzed are daily stock prices of: (i) the TSE from January 1993 to June 2001 and (ii) the S&P 500 index of the NYSE from January 1991 to July 2001. As for the S&P, the daily price data for a different period has been analyzed by Laloux *et al.* [4]. Also, the 30 min price data for the NYSE has been studied by Plerou *et al.* [5,6]. In the TSE data, the number of data points (the days that the market is open) is 1848. We analyze, among all issues in the TSE, the 493 issues which are traded in all of the 1848 days. We select the data of these issues and analyze them. For these data, $N=493$ and $T=1848$. In the S&P 500 data, the number of data points is 2599. We select the issues which have been selected in S&P 500 index before 1991 and analyze their prices. They amount to 297. For these data, $N=297$ and $T=2598$.

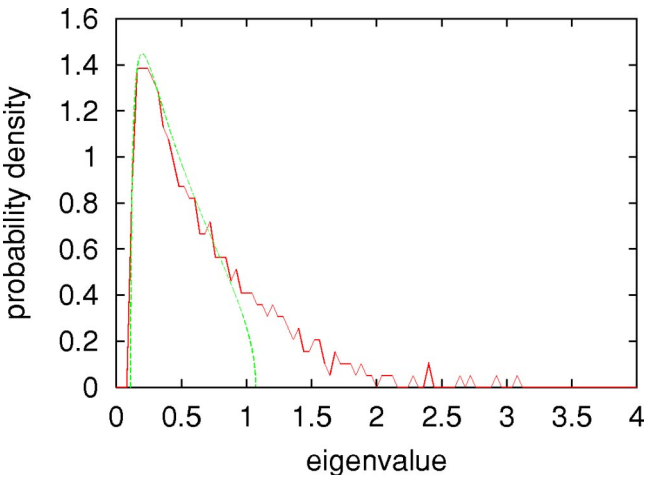

FIG. 1. (Color online) The figure shows the eigenvalue distribution for the correlation matrix of TSE. The line in each figure is for the real data and the dotted line is for the Wishart matrix. We use Eq. (9) multiplied by N'/N for fitting where N' is the number of eigenvalues within $[\lambda_{\min}, \lambda_{\max}]$. σ^2 is fitted to the optimized value by the least-squares method. $\sigma^2 = 0.47(0.53)$ for TSE (S&P). For TSE (S&P), a Kolmogorov–Smirnov test in the fitted region cannot reject the hypothesis that the RMT prediction is the correct description at the 30% (60%) confidence level.

IV. UNIVERSAL RANDOM PROPERTIES OF CROSS CORRELATIONS IN STOCK MARKETS

In Refs. 4 and 5, the cross correlation matrices of the NYSE data are analyzed and found to exhibit remarkable agreement with the predictions of universality properties of RMT for the small eigenvalues' distribution, their nearest- and next-nearest-neighbor spacings, rigidity, and IPR. In this section, we perform a similar analysis on the TSE data and confirm these properties. We also use the S&P data for comparison.

We diagonalize the correlation matrices of TSE and S&P, to obtain the eigenvalues and the eigenvectors $\mathbf{u}_k$ ($k=1,\cdots,N$). k is smaller for a large eigenvalue. For TSE, $\sigma^2=1$ and $Q=N/T=3.75$ give $\lambda_{\min}=0.23$ and $\lambda_{\max}=2.30$, also for S&P, $Q=8.75$ gives $\lambda_{\min}=0.43$ and $\lambda_{\max}=1.79$. We fit the distributions by optimizing σ^2 smaller than 1, as discussed in Ref. 4. Figure 1 shows the eigenvalue distribution for the TSE. We see that the small eigenvalue distribution of the correlation matrix of the TSE is well reproduced by RMT. There are large eigenvalues beyond the bound

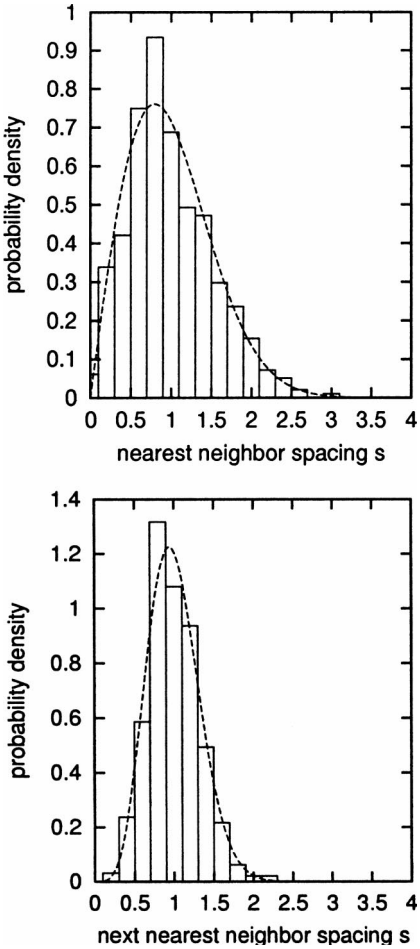

FIG. 2. The figures are the nearest- and the next-nearest-neighbor spacing distribution for TSE compared to the prediction of RMT indicated by the dotted line. A Kolmogorov–Smirnov test cannot reject the hypothesis that the GOE prediction is the correct description at the 30% (80%) confidence level for the nearest-neighbor spacing for TSE (S&P), at the 80% (60%) confidence level for the next-nearest- neighbor spacing for TSE (S&P).

$[\lambda_{\min}, \lambda_{\max}]$ predicted by the Wishart matrix. The largest eigenvalue we obtain is 121.6 (52.2) for the TSE (S&P) and is interpreted as the factor for market trend as readily verified by examining the corresponding eigenvector. The multitude of this factor to the price changes of individual stocks is given by λ_1/N, which is 0.247 (0.176) for the TSE (S&P). Thus, the TSE is more correlated with the trend factor than the S&P.

Next, we compare spacings of the nearest-neighbor and the next-nearest-neighbor eigenvalues, and the rigidity with the predictions of RMT. To examine the statistics of the eigenvalue spacing, we first do the "unfolding" transformation on the data. The unfolding transformation is described in Ref. 6. After doing the unfolding transformation on the eigenvalues below $\lambda_{\max}$, we compare their nearest-neighbor and next-nearest-neighbor spacing distributions to the ones for GOE. The theoretical predictions for the nearest-neighbor spacing and the next-nearest-neighbor spacing are given in Eqs. (17) and (18), respectively. We show in Fig. 2 the spacings of small eigenvalues for the TSE. It shows a good agreement with the prediction of RMT. For the rigidity $\Delta(L)$, the theoretical prediction is given in Eq. (19). The rigidity of the eigenvalues of the cross correlation matrix for the TSE below $\lambda_{\max}$ is compared to RMT in Fig. 3. Figure 3 shows that the rigidity agrees well with the prediction of RMT.

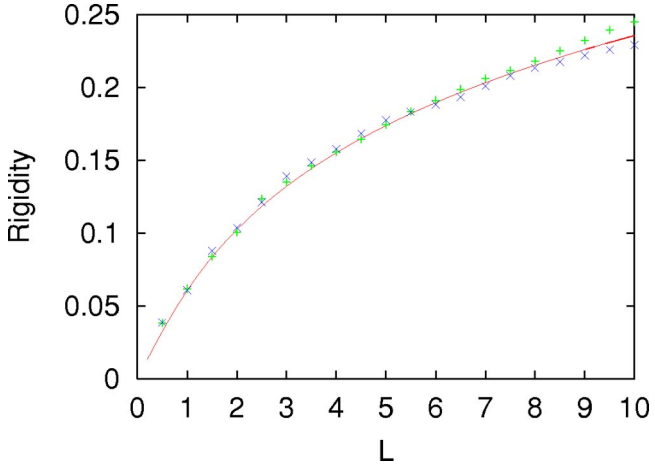

FIG. 3. (Color online) The plus mark is the rigidity $\Delta(L)$ for TSE while the $\times$ mark is the rigidity for S&P. The line is the prediction of RMT. A Kolmogorov–Smirnov test cannot reject the hypothesis that the GOE prediction is the correct description at the 80% confidence level both for TSE and S&P.

In Fig. 4, we plot the calculated IPR for the eigenvectors of the cross correlation matrix of TSE. One sees that the IPR agrees with the prediction of RMT around 1. There are also eigenvectors whose IPRs are larger than the RMT prediction. These eigenvalues are from deterministic correlations. As in Fig. 4, such deviations can be seen at the large eigenvalues. However, one also sees that there is a deviation in small eigenvalues. This deviation is concentrated at the lower edge. A simple model was constructed by Plerou *et al.* [6]. We will study this deviation closely in Sec. VI.

As mentioned, we also performed the same analysis on the S&P data for comparison. Results for the rigidity and IPR are shown in Figs. 3 and 4. We found that the conclusions of Plerou *et al.* [5,6] for 30 min data of the NYSE on eigenvalue spacings also hold for our daily S&P data.

V. STABILITY OF EIGENVALUE DISTRIBUTION OF THE WISHART MATRIX IN THE PRESENCE OF DETERMINISTIC CORRELATIONS

In the previous section, we found that the small eigenvalue distributions of the cross correlation matrices of the TSE and S&P are reproduced well by the ones of the Wishart matrix, as previously found in Ref. 4. The Wishart matrix is generated by the random variables without any deterministic correlations while the real stock data has a distribution of large eigenvalues, showing a deviation from the Wishart matrix. This indicates the existence of deterministic correlations.

Thus, in this section, we examine the stability of the random eigenvalue distribution of the cross correlation matrix W of random variables when one includes deterministic correlations.

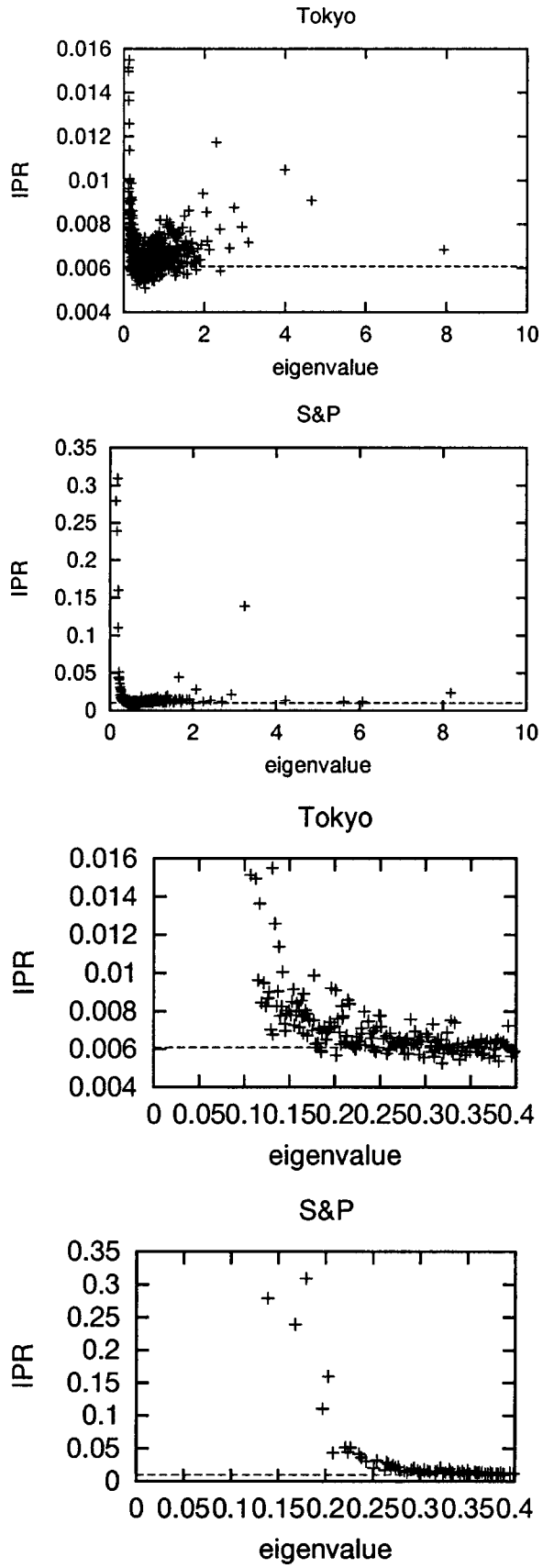

FIG. 4. The upper two figures are IPRs for TSE and S&P. The lower two figures are IPRs for TSE and S&P at small eigenvalues. The dotted lines are the prediction of RMT.

Let us consider a set of random variables for which the deterministic correlation matrix has only a small number of large eigenvalues. We assume that the $T\times N$ matrix $\{M_{ti}=x_i(t)\}$ has a deterministic correlation of the form

$$\langle M_{ti}\rangle = 0, \tag{23}$$

$$\langle M_{ti}M_{\tau j}\rangle = D_{t\tau}\widetilde{C}_{ij}. \tag{24}$$

The cross correlation matrix at step T is given by M^tM. As in RMT, the eigenvalue distribution of M^tM is calculated from the Green function,

$$G(\lambda) \equiv \left\langle \frac{1}{\lambda - M^tM} \right\rangle, \tag{25}$$

by the formula

$$\rho(\lambda) = \frac{1}{2\pi N} \lim_{\epsilon\to 0} \mathrm{Im}[\mathrm{Tr}G(\lambda - i\epsilon) - \mathrm{Tr}G(\lambda + i\varepsilon)]. \tag{26}$$

The present case was studied in Ref. 9. Using the replica method, a Dyson-type equation for G was obtained at $N,T\to\infty$ with $Q=T/N$ fixed as follows:

$$G(\lambda) = \frac{1}{\lambda - \widetilde{C}\mathrm{Tr}\left(\dfrac{D}{1 - D\mathrm{Tr}(\widetilde{C}G(\lambda))}\right)}. \tag{27}$$

Equation (9) is readily obtained by putting $\widetilde{C}=\sigma^2\mathbf{1}$, $D=\mathbf{1}/T$ and taking the trace of Eq. (27)

$$\mathrm{Tr}G(\lambda) = \frac{N}{\lambda - \sigma^2 \dfrac{1}{1 - \dfrac{\sigma^2}{T}\mathrm{Tr}G(\lambda)}}. \tag{28}$$

Solving this second-order algebraic equation for $\mathrm{Tr}G(\lambda)$ and putting the solution to Eq. (26) yields Eqs. (9) and (10).

Now we assume that $\widetilde{C}$ has L large eigenvalues $\lambda_k^{\widetilde{C}}$ $(k=1,2,\cdots,L)$ and the other N-L eigenvalues $\lambda_k^{\widetilde{C}}$ $(k=L+1,\cdots,N)$. We set $\lambda_k^{\widetilde{C}}$ $(k=L+1,\cdots,N)$ to be a same value $\lambda_s^{\widetilde{C}}$. Since the trace of the cross correlation matrix equals N by definition, we have

$$\lambda_s^{\widetilde{C}} = \frac{N - \sum_{k=1}^{L}\lambda_k^{\widetilde{C}}}{N-L}. \tag{29}$$

We also assume no temporal correlations thus set $D=\mathbf{1}/T$.

From Eq. (27), the eigenvalues $\lambda_k^G(\lambda)$ of $G(\lambda)$ are given by

$$\lambda_k^G(\lambda) = \frac{1}{\lambda - \lambda_k^{\widetilde{C}} \dfrac{1}{1 - \dfrac{1}{T}(\mathrm{Tr}_S\widetilde{C}G + \mathrm{Tr}_L\widetilde{C}G)}}. \tag{30}$$

Here, Tr_L and Tr_S are the trace over the eigenspace spanned by the eigenvectors for λ_k^G $(k=1,\cdots L)$, λ_k^G $(k=L+1,\cdots N)$,

respectively. Summation over $k=L+1,\cdots,N$ gives

$$\mathrm{Tr}_S G(\lambda) = \frac{N-L}{\lambda - \lambda_s^{\tilde{C}} \dfrac{1}{1 - \dfrac{1}{T}(\mathrm{Tr}_S \tilde{C} G + \mathrm{Tr}_L \tilde{C} G)}}. \tag{31}$$

For N large, $\rho(\lambda)$ should have finite supports around $\lambda_k^{\tilde{C}}$ in the real axis of λ. We denote supports for large and small eigenvalues D_L and D_S, respectively. We assume the case

$$\lambda_s^{\tilde{C}} \ll \lambda_k^{\tilde{C}}, \quad (k = 1, \ldots L), \tag{32}$$

when D_L and D_S do not have an overlap. In this case, λ_k^G $(k=1,\cdots,L)$ is analytic in D_S while λ_k^G $(k=L+1,\cdots,N)$ has a branch cut. Thus in D_S, $\rho(\lambda)$ is determined by the imaginary part of $\mathrm{Tr}_S G$. For $\mathrm{Tr}_S G$, the contribution from λ_k^G $(k=1,\cdots,L)$ comes from the right-hand side of Eq. (31). Since λ_k^G $(k=1,\cdots L)$ is analytic in the neighborhood of D_S, $\mathrm{Tr}_L G$ is bounded by a constant. As λ_k^G is an algebraic function of N and the scaling behavior consistent with Eq. (30) is $O(1)$, the constant can be taken to be independent of N. Thus, if for $k=1,\cdots,L$

$$L\lambda_k^{\tilde{C}} \ll N\lambda_s^{\tilde{C}}, \tag{33}$$

then

$$\mathrm{Tr}_S \tilde{C} G = \lambda_s^{\tilde{C}} \mathrm{Tr}_S G \gg \mathrm{Tr}_L \tilde{C} G$$

for N large because $\mathrm{Tr}_S G$ gets large as $N \to \infty$. Then, Eq. (31) is approximated by

$$\mathrm{Tr}_S G(\lambda) = \frac{N-L}{\lambda - \lambda_s^{\tilde{C}} \dfrac{1}{1 - \dfrac{\lambda_s^{\tilde{C}}}{T} \mathrm{Tr}_S G(\lambda)}}. \tag{34}$$

Equation (34) is equal to Eq. (28) when $\sigma^2 = \lambda_s^{\tilde{C}}$ and N is replaced by N-L. By putting the solution of Eq. (34) to Eq. (26), we get

$$\rho(\lambda) \simeq \frac{N-L}{N} \frac{Q}{2\pi\lambda_s^{\tilde{C}}} \frac{\sqrt{(\lambda_{\max} - \lambda)(\lambda - \lambda_{\min})}}{\lambda}. \tag{35}$$

This formula is valid under Eqs. (32) and (33). Note that there is a trade off between $N, L, \lambda_s^{\tilde{C}}, \lambda_k^{\tilde{C}} (k=1,\cdots,L)$ under Eqs. (32) and (33). Thus, the N-L eigenvalue distribution of this model can be approximated by the one for the Wishart matrix.

To conclude, the distribution of the small eigenvalues remains the same in the $N \to \infty$, as long as the numbers of the large eigenvalues of the deterministic correlation $\tilde{C}$ are finite and they appear only outside of D_S.

To confirm the validity of the approximation, we performed a Monte Carlo simulation with six large eigenvalues. We choose the large eigenvalues to be 121.6, 14.5, 11.4, 7.9, 4.7, and 4.0 which are the observed large eigenvalues of the TSE. The result is shown in Fig. 5. We see that the large eigenvalues correspond to the large eigenvalues of the real correlation matrix while the small eigenvalue distribution is well reproduced by the one for the Wishart matrix. We also examined other values of large eigenvalues and obtained similar results. Moreover, the probability of observed eigenvalues has a finite width by the effect of randomness. The width of an observed eigenvalue is wider for a larger eigenvalue.

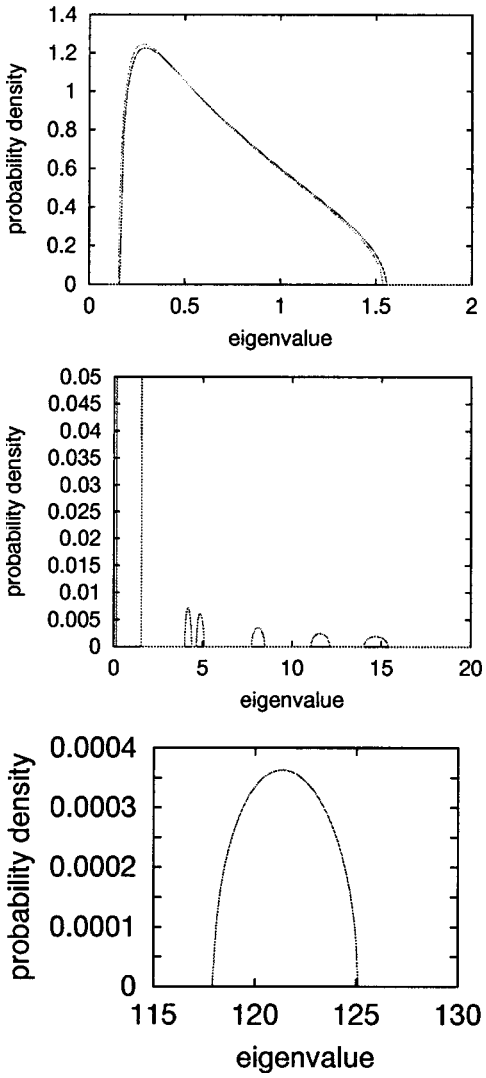

FIG. 5. (Color online) The line is for the model with large eigenvalues of the real correlation matrix while the dotted line is for Eq. (9) with $\sigma^2 = \lambda_s^{\tilde{C}}$ in Eq. (29). The small eigenvalue distribution (the upper graph) is very close. The middle graphs are the large eigenvalue distribution. We take the large eigenvalues of the real correlation matrix as 121.6, 14.5, 11.4, 7.9, 4.7, and 4.0 which are found for TSE. The eigenvalues are observed in the neighborhood of these values. Also, the observed eigenvalues have a finite width by the effect of randomness. The width of the observed eigenvalues is wider for the larger eigenvalues.

VI. LEVEL REPULSION OF DETERMINISTIC CORRELATIONS BY RANDOMNESS

According to Plerou *et al.* [5], the deviation at small eigenvalues arises from strong correlations among a small

number of issues. This is illustrated well by the following model. We consider a model that N issues have an equal correlation c:

$$\tilde{C} = \begin{pmatrix} 1 & c & \cdots & c \\ c & 1 & & \vdots \\ \vdots & & \ddots & c \\ c & \cdots & c & 1 \end{pmatrix}, \tag{36}$$

where $\tilde{C}$ has an eigenvalue $1+(N-1)c$ with no degeneracy and an eigenvalue $1-c$ with degeneracy $N-1$. The eigenvalue $1-c$ becomes small if c is close to 1, i.e., strong correlation. Its eigenvectors have nonzero components at the correlated issues, resulting in a large IPR.

However, this reasoning of large IPR eigenvectors at small eigenvalues is not sufficient to explain two facts. First, eigenvectors with a large IPR appear only *below* the bulk of the eigenvalue distribution of the Wishart matrix, concentrating at the lower edge. Since the correlation c should be distributed in a wide range, eigenvectors with a large IPR should also be distributed in a wide range. Thus, the absence of small eigenvalues with a large IPR within the bulk is puzzling. Second, each eigenvector with a large IPR is observed at a smaller value than expected from the model above. As the largest nondiagonal element of the correlation matrix of the TSE (S&P) is 0.74 (0.83), Eq. (36) tells us that the eigenvector with a large IPR and with the smallest eigenvalue should be observed at 0.26 (0.18). Actually, the smallest eigenvalue with a large IPR is observed at 0.11 (0.14) which is smaller than the lower bound of the eigenvalue distribution of the Wishart matrix.

These two facts motivate us to study the interplay between deterministic correlations and randomness. We consider a model of random variables with a deterministic correlation matrix $\tilde{C}$, and examine the IPRs of eigenvectors of the cross correlation matrix C. As a simple model, we assume $\tilde{C}$ to have a following form:

$$\tilde{C} = \begin{pmatrix} \tilde{C}_1 & 0 & \cdots & \cdots & 0 \\ 0 & \tilde{C}_2 & & & \vdots \\ \vdots & & \ddots & & \\ \vdots & & & \tilde{C}_L & 0 \\ 0 & \cdots & \cdots & 0 & \mathbf{1} \end{pmatrix}. \tag{37}$$

Here, $\tilde{C}_l$ $(l=1,\cdots,L)$ and $\mathbf{1}$ are

$$\tilde{C}_l = \begin{pmatrix} 1 & c_l & \cdots & c_l \\ c_l & 1 & & \vdots \\ \vdots & & \ddots & c_l \\ c_l & \cdots & c_l & 1 \end{pmatrix} \qquad \mathbf{1} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix}. \tag{38}$$

The form of $\tilde{C}$ assumes L groups of issues with strong correlations. We consider N random variables $x_i(t)$ with

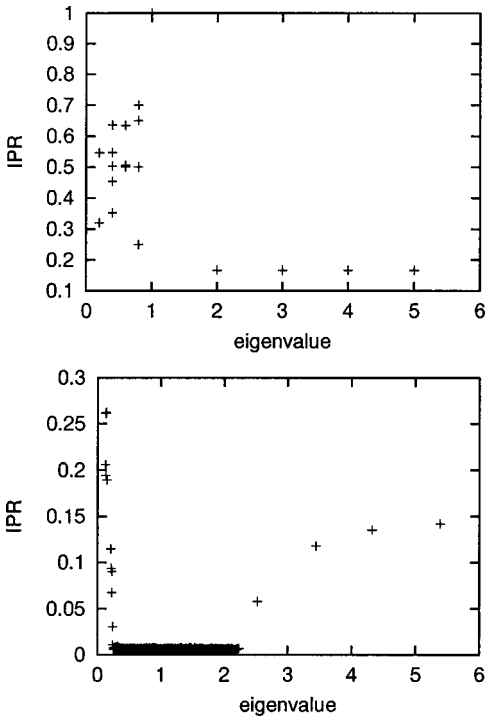

FIG. 6. The upper graph is the IPR of the eigenvectors of the real correlation matrix $\tilde{C}$ given by Eqs. (37)–(40). The lower graph is the IPR for the eigenvector of C. In the simulation, we set $N=493$, $T=1847$, $M=6$, $L=4$, $c_1=0.8$, $c_2=0.6$, $c_3=0.4$, and $c_4=0.2$.

$$\langle x_i(t)x_j(\tau)\rangle = \tilde{C}_{ij}\delta_{t\tau}, \tag{39}$$

and examine their T-step cross correlation matrix

$$C_{ij} = \frac{1}{T}M^tM = \frac{1}{T}\sum_{t=1}^{T} x_i(t)x_j(t). \tag{40}$$

We set $N=493$ and $T=1847$ following our TSE data. We set the number L of strongly correlated groups to be 4 and the number of issues M participating each group to be 6. We choose the correlations to be $c_1=0.8$, $c_2=0.6$, $c_3=0.4$, and $c_4=0.2$. We performed a Monte Carlo simulation of this model. We present an IPR of the eigenvalues in Fig. 6. Figure 6 shows that eigenvalues with a large IPR distribute outside the bounds of eigenvalue distribution from randomness as in the real stock data. In this model, there should be 20 (counting degeneracies) small eigenvalues with a large IPR in the simple model above, but the observed ones with a large IPR only amount to 10. This implies that, *when small eigenvalues arising from a strong correlation appear within the bounds of the Wishart matrix, IPRs of their eigenvectors get smaller and cannot be distinguished from the random eigenvalues*. This is one effect of randomness on deterministic correlations. We also note that even for the eigenvectors which have a larger IPR than the RMT value, their IPRs are smaller than expected.

Moreover, $\tilde{C}$ has small eigenvalues 0.2 and 0.4 while the corresponding eigenvalues of C distribute in the vicinity of 0.14 and 0.22, respectively. On the other hand, the eigenval-

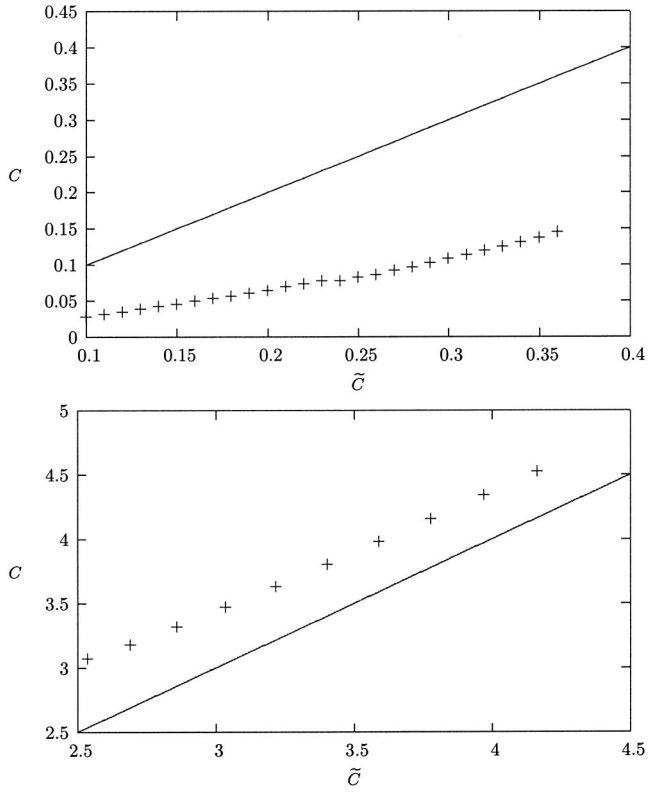

FIG. 7. The effect of level repulsion on the eigenvalues of C. The horizontal axis is the small (large) eigenvalue of $\tilde{C}$ and the vertical axis is the corresponding eigenvalue of C. The upper (lower) graph is for the case where eigenvalues of $\tilde{C}$ are smaller (larger) than 1. The crosses are the result of a Monte Carlo simulation based on Eqs. (39) and (40). The straight line corresponds to the absence of the effect of randomness, when the eigenvalues of C are identical to those of $\tilde{C}$. The eigenvalues of C are repelled from the bulk vicinity of 1.

ues of C corresponding to the large eigenvalues of $\tilde{C}$ are shifted to values larger than the original ones. Namely, *the eigenvalues of C from the deterministic correlation are repelled from the distribution of the random eigenvalues*. We performed Monte Carlo simulations by changing the parameters for $\tilde{C}$ and got similar results. This may be interpreted as a manifestation of the universal effect of randomness, called "level repulsion" [13]. According to RMT, the eigenvalues of random matrices are repelled from each other by the logarithmic potential in $-\ln|\lambda_i-\lambda_j|$ in Eq. (13). Even when some deterministic terms are present, this logarithmic potential causes a repulsion between eigenvalues. This universal effect has been observed for various systems such as levels of complicated nuclei. In the present case, deterministic correlations between random variables are repelled from the bulk distribution of the random eigenvalues. The eigenvalues in the RMT bounds form a repulsive potential and it repels the eigenvalues outside them.

We can deduce this level repulsion by solving the Dyson-type equations (23)–(26) numerically. We assume for simplicity that the eigenvalues of $\tilde{C}$ are 1 except one eigenvalue smaller or larger than 1. We solve Eqs. (23)–(26) numerically

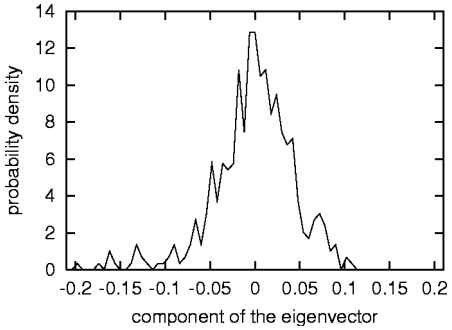

FIG. 8. The component distribution of the eigenvector for the sixth largest eigenvalue 4.0 of TSE. The components distribute continuously and it is hard to distinguish the components from correlations.

for $N=293$ and $T=1847$ and obtain the relation between the smaller (or larger) eigenvalue and the corresponding eigenvalue of C. The result is shown in Fig. 7. Figure 7 shows that smaller (larger) eigenvalues of $\tilde{C}$ are repelled by the bulk distribution around 1 and are observed as smaller (larger) eigenvalues of C.

Thus, we found two interplays between deterministic correlations and randomness. Namely, when groups of issues have strong correlations, it results in large and small eigenvalues in the cross correlation matrix. Some of these eigenvalues are soaked up within the RMT bounds and their IPRs becomes as small as the RMT value. They cannot be distinguished from random eigenvalues. On the other hand, eigenvalues from deterministic correlations outside the RMT bounds feel the repulsive potential generated by the bulk distribution of randomness. At the lower edge, they are shifted to smaller values. We believe that these give the explanation for two deviations we raised in this section.

VII. GROUPS OF ISSUES FORMED BY STRONG CORRELATION

We have seen that the existence of a group of issues with strong correlation results in eigenvalues of the cross correlation matrix with a large IPR. Conversely, by examining the eigenvectors with large IPR, we may identify groups formed by strong correlations.

For the NYSE, Plerou *et al.* [5] examined the eigenvectors of large eigenvalues and distinguished strongly correlated issues by a criteria to have a large component in these eigenvectors. They found that the groups are formed according to the industrial sectors. However, we found a difficulty in applying their method to the TSE. Because eigenvectors for large eigenvalues have significant components not only from correlations but also from randomness, even if an issue has a large component in an eigenvector of large eigenvalue, it is difficult to tell whether it is from the effect of deterministic correlation or just from randomness. Especially for the TSE, the effect of deterministic correlations is apparently not strong enough to make the separation straightforward. As we examined the eigenvectors of the large eigenvalues, we found it impossible to separate the group of strongly correlated issues. For example, Fig. 8 shows the component dis-

TABLE I. TSE issues with $Z_i \geqslant \alpha_{\mathrm{th}}$.

Eigenvector	TSE code	Company Name	Sector
u_2	6701	NEC	Electric products
u_2	6702	Fujitsu	Electric products
u_2	8035	Tokyo Electron	Electric products
u_3	1888	Wakachiku Construction	Construction
u_3	8834	Douwa Real Estate	Real estate
u_4	9501	Tokyo Electric Power	Electric power
u_4	9503	Kansai Electric Power	Electric power
u_4	9504	Chuugoku Electric Power	Electric power
u_4	9506	Tohoku Electric Power	Electric power
u_4	9509	Hokkaido Electric Power	Electric power
u_5	1888	Wakachiku Construction	Construction
u_5	8834	Douwa Real Estate	Real estate
u_5	1801	Taisei Corporation	Construction
u_5	1804	Satou Kogyo	Construction
u_5	1805	Tobishima Construction	Construction
u_5	1806	Fujita Corporation	Construction
u_5	1886	Aoki Corporation	Construction
u_5	8601	Daiwa Securities	Finance
u_5	8603	Nikko Cordial Group	Finance
u_6	8834	Douwa Real Estate	Real estate
u_6	9501	Tokyo Electric Power	Electric power
u_6	9503	Kansai Electric Power	Electric power
u_6	9504	Chuugoku Electric Power	Electric power
u_6	9506	Tohoku Electric Power	Electric power
u_6	9509	Hokkaido Electric Power	Electric power
u_6	1804	Sato Corporation	Construction
u_6	1805	Tobishima Construction	Construction
u_6	1806	Fujita	Construction
u_6	1886	Aoki Corporation	Construction
u_7	9504	Chuugoku Electric Power	Electric power
u_7	9506	Tohoku Electric Power	Electric power
u_7	5801	Furukawa Electric	Nonferrous metal
u_7	8004	Nichimen	Wholesale
u_8	8335	Ashikaga Bank	Bank
u_8	9766	Konami	Service
u_9	8004	Nichimen	Wholesale
u_9	8335	Ashikaga Bank	Bank
u_9	8752	Sumitomo Mitisui Kaijyo	Insurance

tribution of the eigenvector for the sixth largest eigenvalue 4.0 in the TSE. One sees that the components have a continuous distribution and it is hard to separate large components due to deterministic correlations.

Therefore, here we propose a supplementary method to identify strongly correlated components. As we saw in Sec. VI, when a group of issues is formed by strong correlations, they not only have a large component in the eigenvectors of the corresponding large eigenvalue, but also have a large component in the eigenvectors of the corresponding small eigenvalue. On the other hand, issues which do not have strong correlations with others should have the normal distribution in eigenvectors. Namely, the deviation from the normal distribution indicates that the issue is correlated with others. To quantify how an issue has a distribution different from the normal distribution, we define a quantity Z_i as follows:

$$Z_i = \sum_{k:I_k \geqslant \delta_{\mathrm{th}}} u_{ki}^2, \tag{41}$$

where δ_{th} is a threshold for an IPR. Z_i is the sum of the square of ith component of the eigenvectors which have an

TABLE II. S&P issues with $Z_i \geqslant \alpha_{\mathrm{th}}$.

Eigenvector	Ticker	Company Name	Industries
u_2	AEP	American Electric Power	Electric power
u_2	DUK	Duke Energy Corporation	Electric power, natural gas
u_3	APC	Anadarko Perroleum Corp.	Oil,gas
u_3	BHI	Baker Hughes Inc.	Oil, related
u_3	XoM	Exxon Mobil Corporation	Oil, coal, copper
u_3	HAL	Halliburton Company	Oil, gas
u_3	RD	Royal Dutch Petroleum Co.	Oil, gas, chemical
u_3	SLB	Schlumberger Ltd.	Oil
u_3	UCL	Unocal Corporation	Oil, gas
u_4	GP	Georgia-Pacific Group	Paper manufacturer, pulp
u_4	IP	International Paper Co.	Paper manufacturer
u_4	MEA	Mead Corporation	Paper Manufacturer pulp, gum
u_4	WY	Weyerhaeuser Company	Paper Manufacturer, pulp, forestry, wooden goods
u_5	MRK	Merck & Co., Inc.	Medicine manufacturer
u_5	PFE	Pfizer Inc.	Medicine manufacturer
u_5	SGP	Schering-Plough Corp.	Medicine Manufacturer
u_6	BK	Bank of New York Co.	Bank
u_6	JPM	J.P. Morgan Chase & Co.	Finance
u_6	PNC	PNC Financial Services	Finance
u_6	STI	SunTrust Banks, Inc.	Bank
u_7	ABX	Barrick Gold Corp.	Gold mining, gold goods
u_7	HM	Homestake Mining Co.	Gold mining
u_7	NEM	Newmont Mining Corp.	Gold mining
u_7	PDG	Placer Dome Inc.	Gold mining
u_8	SBC	SBC Communications Inc.	Telecommunication, cable television, internet
u_8	VZ	Verizon Communications	Telecommunication, internet
u_8	MU	Micron Technology, Inc.	Semiconductor
u_8	TXN	Texas Instruments	Semiconductor
u_9	AMR	AMR Corporation	Aviation
u_9	DAL	Delta Air Lines, Inc.	Aviation
u_9	F	Ford Motor Company	Automobile
u_9	GM	General Motors Corp.	Automobile
u_{10}	EIX	Edison International	Holding company of electric power
u_{10}	PCG	PG&E Corporation	Holding company of electric power
u_{11}	AL	Alcan Inc.	Aluminium, aluminium can
u_{11}	AA	Alcoa, Inc.	Aluminium

IPR $\geqslant \delta_{\mathrm{th}}$. We set $\delta_{\mathrm{th}}=0.008(0.02)$ for the TSE (S&P), which sort out 41(28) eigenvectors. If the ith issue has no true correlation with others, the components u_{ki} of the eigenvectors follow the normal distribution, and hence the probability of having a large Z_i should be small. Thus, the ith issue may be regarded as significantly correlated if Z_i is larger than a certain threshold α_{th}. We choose α_{th} so that the probability of $Z_i \geqslant \alpha_{\mathrm{th}}$ is 1% if the eigenvector components for the ith issue follow the normal distribution. For our data, $\alpha_{\mathrm{th}}=0.131$ (0.162) for the TSE (S&P). If the ith issue has a large component in an eigenvector, we consider it to be in the corresponding group of the strong correlations when $Z_i \geqslant \alpha_{\mathrm{th}}$.

We applied this method to large eigenvalues observed in our market data. The results are shown in Tables I and II.

In the S&P, the electric power sector, and oil and gas related sectors play major parts in the correlations. In the TSE, the electric products sector and construction sector play major parts.

In the S&P, each eigenvector corresponds to an industrial sector. This means that each industrial sector forms a strongly correlated group. On the other hand, in the TSE, there are eigenvectors whose participants are from different

industrial sectors, which may indicate a more complicated correlation structure of the market. Thus it seems that the TSE and S&P (NYSE) have some differences in the structure of the correlations, while the "random" part is well described by the universal theory in the both markets. It would be interesting to find the origin of the difference. This might be useful to give some insights into the difference of the economic structures of the two countries.

As far as our data samples are concerned, we may conclude that the method which we propose utilizing small eigenvectors and their IPR effectively distinguishes strongly correlated groups in the markets.

We noticed that Giada *et al.* investigated the grouping of S&P data in Ref. 14 based on a model considered by Noh [15]. The method proposed in Refs. 14 and 15 has the advantage of directly giving the "noise-undressed" correlation matrix. However, the basic assumption of their method is that each issue belongs to only one cluster of correlated issues. This assumption is apparently not quite true according to our analysis. For example, "Tohoku Electric Power" appears in three different groups in Table I. Therefore, we believe that more analysis based on conservative assumptions should be made before applying the estimated true correlation to the portfolio management.

VIII. CONCLUSIONS

We analyzed the eigenvalues and the eigenvectors of the cross correlation matrices of the TSE and NYSE (S&P500) stock market data. We found that results of Refs. 4–6 reported for the NYSE are also valid for the TSE. The eigenvalue distribution obeys the RMT prediction in the bulk but there are some deviations at the large eigenvalues. We also examined the nearest-neighbor spacing, the next-nearest-neighbor spacing, and the rigidity of the eigenvalues and found that they follow the universality of GOE. These are consistent with Refs. 4–6 and imply that the large eigenvalues are due to the existence of correlations while the eigenvalues distributed in the bulk are due to randomness. We also examined the IPRs of the eigenvectors of the correlation matrices. In the bulk, the IPR distribution follows the prediction of GOE, but there are deviations outside the RMT bounds. Plerou *et al.* [5,6] argued that deviations at the lower edge are due to strong correlations. We found that this reasoning is qualitatively valid, but quantitatively it cannot explain the fact that small eigenvalues with a large IPR concentrate at the lower edge and the observed eigenvalues are smaller than the expected values.

To explain this phenomenon, we studied RMT with deterministic correlations. We found that each eigenvalue from deterministic correlations is observed at values which are repelled from the bulk distribution. We interpreted this repulsion as a reminiscent of the effect of randomness, known as level repulsion. This effect is shown to be deduced by solving the Dyson-type equation numerically.

We also proposed a method to distinguish strongly correlated groups of issues based on the IPR. It reduces the accidental appearance of uncorrelated issues. Applying this method, we found that issues of the S&P are grouped according to the industrial sectors. On the other hand, issues of the TSE are grouped in more complicated ways, suggesting some differences in the structure of the markets.

ACKNOWLEDGMENTS

The authors acknowledge the Institute for Asset Management of Mizuho T. B. and Hiraku Kusaka at BNP Paribas for providing the stock price data. They thank Hiraku Kusaka also for his critical reading of the manuscript and useful comments, and Shinobu Hikami for discussions.

[1] R. C. Merton, Bell J. Econo. and Man. Sci. **4**, 141 (1973); F. Black and M. Scholes, J. Political Econo. **81**, 637 (1974).
[2] E. J. Elton and M. J. Gruber, *Modern Portfolio Theory and Investment Analysis* (Wiley, New York, 1995); H. Markowitz, *Portfolio Selection: Efficient Diversification of Investments* (Wiley, New York, 1959).
[3] See e.g., R. N. Mantegna and H. E. Stanley, *An Introduction to Econophysics* (Cambridge University Press, Cambridge, UK, 2000); J. P. Bouchaud and M. Potters, *Theory of Financial Risk* (Cambridge University Press, Cambridge, UK, 2000).
[4] L. Laloux, P. Cizeau, J.-P. Bouchaud, and M. Potters, Phys. Rev. Lett. **83**, 1467, (1999).
[5] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. N. Amaral, and H. E. Stanley, Phys. Rev. Lett. **83**, 1471 (1999).
[6] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. N. Amaral, T. Guhr, and H. E. Stanley, Phys. Rev. E **65**, 066126 (2002).
[7] T. H. Baker, P. J. Forrester, and P. A. Pearce, J. Phys. A **31**, 6087 (1998).
[8] A. Edelman, SIAM J. Matrix Anal. Appl. **9**, 543 (1998).
[9] A. M. Sengupta and P. P. Mitra, Phys. Rev. E **60**, 3389 (1999).
[10] S. Hikami, Prog. Theor. Phys. **92**, 479 (1994).
[11] M. L. Metha, *Random Matrices* (Academic Press, New York, 1991).
[12] Y. V. Fyodorov and A. D. Mirlin, Phys. Rev. Lett. **69**, 1093 (1992); **71**, 412 (1993); A. D. Mirlin and Y. V. Fyodorov, J. Phys. A **26**, L551 (1993).
[13] C. E. Potter and N. Rosenzweig, Phys. Rev. **120**, 1698 (1956).
[14] L. Giada and M. Marsili, Phys. Rev. E **63**, 061101 (2001).
[15] J. D. Noh, Phys. Rev. E **61**, 5981 (2000).