Systematic analysis of group identification in stock markets

Dong-Hee Kim* and Hawoong Jeong†
Department of Physics, Korea Advanced Institute of Science and Technology, Daejeon 305-701, Korea
(Received 8 March 2005; published 25 October 2005)

We propose improved methods to identify stock groups using the correlation matrix of stock price changes. By filtering out the marketwide effect and the random noise, we construct the correlation matrix of stock groups in which nontrivial high correlations between stocks are found. Using the filtered correlation matrix, we successfully identify the multiple stock groups without any extra knowledge of the stocks by the optimization of the matrix representation and the percolation approach to the correlation-based network of stocks. These methods drastically reduce the ambiguities while finding stock groups using the eigenvectors of the correlation matrix.

DOI: 10.1103/PhysRevE.72.046133

PACS number(s): 89.65.Gh, 05.40.Ca, 89.75.Fb, 89.75.–k

I. INTRODUCTION

The study of correlations in stock markets has attracted much interest of physicists because of its challenging complexity as a complex system and its possible future applications to the real markets [1]. In the early years, a correlation-based taxonomy of stocks and stock market indices was studied by the method of the hierarchical tree [2,3]. Recently, the minimum spanning tree technique was introduced to study the structure and dynamics of the stock network [4–6], the random matrix theory was applied to find out the difference between the random and nonrandom property of the correlations [7–11], and the maximum likelihood clustering method was developed and applied to identify cluster structures in stock markets [12]. Also, these studies have been extended to the applications to the portfolio optimization in real market [5,9,13].

Commonly, the correlation between stocks is expressed by the Pearson correlation coefficient of log-returns,

$$G_{ij}(t) = \ln S_{i}(t + \Delta t) - \ln S_{j}(t),$$

where $S_{i}(t)$ is the price of stock $i$ at time $t$. From real time series data of $N$ stock prices, we can calculate the element of $N \times N$ correlation matrix $C$ as follows

$$C_{ij} = \frac{[G_{i}(t) - \langle G_{i} \rangle][G_{j}(t) - \langle G_{j} \rangle]}{\sqrt{[G_{i}^{2} - \langle G_{i}^{2} \rangle] [G_{j}^{2} - \langle G_{j}^{2} \rangle]}},$$

where $\langle \cdots \rangle$ indicates time averages over the period of the time series. By definition, $C_{ii} = 1$ and $C_{ij}$ has a value in $[-1, 1]$.

Laloux et al. [7] and Plerou et al. [8,9] studied the statistical properties of an empirical correlation matrix between stock price changes defined in Eq. (2) for real markets. In comparison with the prediction of the random matrix theory, they found that the statistics of the bulk eigenvalues are in remarkable agreements with the universal properties of the random correlation matrix. For example, the bulk part of the eigenvalue spectrum of the empirical correlation matrix for $N$ stocks over $L$ price data has the form of the spectrum of the random correlation matrix [14] which is given by

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

for $\lambda \in [\lambda_{\min}, \lambda_{\max}]$ in the limit of $N, L \to \infty$ with fixed $Q \equiv L/N$, where $\lambda_{\max} = (1 + 1/\sqrt{Q})^2$ and $\lambda_{\min} = (1 - 1/\sqrt{Q})^2$. Moreover, the level spacing statistics of eigenvalues exhibits good agreement with the results from the Gaussian orthogonal ensemble of random matrices [8,9].

On the other hand, the nonrandom properties of the correlation matrix have also been studied with the empirical correlation matrix [8–10]. From the empirical data for the New York Stock Exchange, it was found that each eigenvector, corresponding to the few largest eigenvalues larger than the upper bound of the bulk eigenvalue spectrum, is localized, in a sense that only a few components contribute to the eigenvector mostly, and the stocks corresponding to those dominant components of the eigenvector are found to belong to a common industry sector. Very recently, Utsugi et al. confirmed and improved those results through the similar analysis for the Tokyo Stock Exchange [11].

In order to confirm the localization property of eigenvectors, we perform the similar analysis to the previous studies [8–10] on eigenvectors of the correlation matrix using our own dataset of stock prices. We analyze the daily prices of $N=135$ stocks belonging to the New York Stock Exchange (NYSE) for the 20-year period 1983–2003 ($L=5000$ trading days) which is publicly available from the web-site (http://finance.yahoo.com) [15]. Indeed, if we put stocks in the order of their industrial sectors, we observe that the eigenvector components corresponding to stocks which belong to specific industrial sectors give high contributions to each of the eigenvectors for the few largest eigenvalues (see Fig. 1). For instance, the stocks belonging to the energy, technology, transportation, and utilities sectors highly contribute to the eigenvector for the second largest eigenvalue; the energy sector constitutes the big part of the eigenvector for the third largest eigenvalue; the fourth largest eigenvalue gives the eigenvector localized on the basic materials, consumer (non-
corresponding to the second to fifth largest eigenvalues \( \lambda_2 - \lambda_4 \) of the correlation matrix. The stocks are sorted by industrial sectors, A: basic materials, B: capital goods, C: conglomerates, D: consumer (cyclical), E: consumer (noncyclical), F: energy, G: financial, H: healthcare, I: services, J: technology, K: transportation, and L: utilities, which are separated by dashed lines.

cyclical), healthcare, and utilities sectors; the eigenvector for the fifth largest eigenvalue is also localized on several specific industrial sectors.

However, it is not straightforward to find out specific stock groups, such as the industrial sectors, inversely. If each of the eigenvectors had well-defined dominant components and the corresponding set of stocks were independent of the sets from other eigenvectors, it would become easy to identify the stock groups. Unfortunately, in our study, it turns out that not only the set of eigenvector components with dominant contribution can be hardly defined in the eigenvector but also such a set is likely to overlap with the sets from other eigenvectors unless we pick a very small number of stocks with few highest ranks of their contributions to the eigenvectors; Figure 1 indicates that each of the eigenvectors is localized on a multiple number of industrial sectors and the corresponding stocks severely overlap with those from the other eigenvectors. Therefore it is very ambiguous to identify the stock groups for practical purposes. The aim of this study is to get rid of these ambiguities and finally find out relevant stock groups without any aid of the table of industrial sectors.

In this paper, we introduce the improved method to identify stock groups which drastically reduce the ambiguities in finding multiple groups using eigenvectors of the correlation matrix. We first filter out the random noise and the market-wide effect from the correlation matrix. With the filtered correlation matrix, we apply optimization and percolation approaches to find the stock groups. Through the optimization of the stock sequences representing the matrix indices, the filtered correlation matrix is transformed into the block diagonal matrix in which all stocks in a block are found to belong to the same group. By constructing a network of stocks using the percolation approach on the filtered correlation matrix, we also successfully identify the stock groups which appear in the form of isolated clusters in the resulting network.

This paper is organized as follows. In Sec. II, the detailed filtering method to construct the group correlation matrix is given. For the filtering, the largest eigenvalue and the corresponding eigenvector are required and they are calculated from the first-order perturbation theory. In Sec. III entailed stock group finding methods using the optimization and the percolation are given and the resulting stock groups are specified. In Sec. IV, a summary and conclusions are presented.

II. GROUP CORRELATION MATRIX

A. Filtering

The group of stocks is defined as a set of highly interrelated stocks in their price changes. In the empirical correlation matrix, because several types of noises are expected to coexist with the intragroup correlations, it is essential to filter out such noises to isolate the intragroup correlations which we are interested in. With the complete set of eigenvalues and eigenvectors, the correlation matrix in Eq. (2) can be expanded as

\[
C = \sum_{\alpha = 0}^{N-1} \lambda_\alpha |\alpha\rangle \langle \alpha|,
\]

where \( \lambda_\alpha \) is the eigenvalue sorted in descending order and \( |\alpha\rangle \) is the corresponding eigenvector. Because only the eigenvectors corresponding to the few largest eigenvalues are believed to contain the information on significant stock groups, we can identify a filtered correlation matrix for stock groups by choosing a partial sum of \( \lambda_\alpha |\alpha\rangle \langle \alpha| \) relevant to stock groups, which we will call the group correlation matrix, \( C^g \).

In order to extract \( C^g \) from the correlation matrix, taking the previous results of Pléou et al. [8–10] for granted, we posit that the eigenvalue spectrum of the correlation matrix is organized by the marketwide part of the largest eigenvalue, the group part of intermediate discrete eigenvalues, and the random part of small bulk eigenvalues. Then, we can separate the correlation matrix into three parts as

\[
C = C^m + C^g + C^r = \lambda_0 |0\rangle \langle 0| + \sum_{\alpha = 1}^{N_0} \lambda_\alpha |\alpha\rangle \langle \alpha| + \sum_{\alpha = N_0 + 1}^{N-1} \lambda_\alpha |\alpha\rangle \langle \alpha|,
\]

where \( C^m \), \( C^g \), and \( C^r \) indicate the marketwide effect, the group correlation matrix, and the random noise terms, respectively.

While the determination of \( C^m \) is straightforward, it is not so clear to determine \( N_0 \) for separating \( C^g \) and \( C^r \). If there were no correlation between stock prices, the bulk eigenvalues have to follow Eq. (3), and thus the upper bound of the bulk eigenvalues can be clearly determined from \( Q \). How-
This decomposition of the correlation matrix gives non-trivial characteristics to the distribution of the group correlation matrix elements $C^g_{ij}$. In Fig. 2(c), it turns out that the distribution of $C^g_{ij}$ shows positive heavy tail. This indicates that $C^g$ contains a non-negligible number of strongly correlated stock pairs, which is expected to come from the correlation between the stocks belonging to the same group. On the other hand, $C^r$ shows the Gaussian distribution consistent with the prediction of the random matrix theory [9]. While this Gaussian-like distribution is also observed partially in the distribution of $C^c_{ij}$ due to the coupling between group correlations and random noises, it turns out that this remaining noise does not seriously affect the identification of stock groups. The distribution of $C^c_{ij}$ shows that $C^m$ also contains highly correlated stock pairs, but we find that $C^m_{ij}$ is not relevant to the group correlation and thus have to be filtered out for the clear identification of the stock groups, which is discussed in Sec. II B.

Since the quality of the correlation matrix can depend on the period of empirical data or generally $Q=L/N$, our decomposition of the correlation matrix can also depend on $Q$. Here we simply check how the determination of $N^c$ and the resulting matrix element distribution of the decomposed matrices are changed depending on $Q$ (see Fig. 3). For $Q=19.4$ (1993–2003) and $Q=7.0$ (1999–2003), $C^g$ and $C^r$ are separated at $N^c=10$, which is not very different from $N^v=9$ of the larger dataset we use throughout this paper, and in addition, the distribution of the matrix element shows the similar degree of the heavy tail in $C^g_{ij}$. However, decreasing $Q$ much smaller, the bulk eigenvalue spectrum becomes wider so that more eigenvalues relevant to the group correlation can be buried in the bulk spectrum, which leads to smaller $N^c$ that turns to be 7 for $Q=1.7$ (2003). Even in this case of $Q=1.7$, the positive heavy tail is still found in $C^g_{ij}$ but very weaker than higher $Q$’s. These imply that we need a large enough $Q$ for the stock group identification.

**B. Largest eigenvalue and corresponding eigenvector**

Our filtering is based on the following interpretations of the previous studies: the bulk part of the eigenvalues and
their eigenvectors are expected to show the universal properties of the random matrix theory and the largest eigenvalue and its eigenvector are considered as a collective response of the entire market [8–10]. While the random characteristics of the bulk eigenvalues have been studied intensively, only the empirical tests have been done for the largest eigenvalue and its eigenvector so far [9,10]. Thus, to understand the more accurate meaning, we calculate the largest eigenvalue and its eigenvector of the correlation matrix by using perturbation theory [17,18].

In stock markets, it has been understood that there exist three kinds of fluctuations in stock price changes: a marketwide fluctuation, synchronized fluctuations of stock groups, and a random fluctuation [8–10]. For simplicity, we consider a situation in which a system with only the marketwide fluctuation is perturbed by other fluctuations. Let us assume that the price changes of all the stocks in the market find a synchronized background fluctuation with zero mean and variance $C_{0}$ as a marketwide effect. Then, we can write down the $N \times N$ unperturbed correlation matrix as

$$
C^{0} = \begin{pmatrix}
1 & c_{0} & \cdots & c_{0} \\
c_{0} & 1 & \ddots & \vdots \\
\vdots & \ddots & \ddots & c_{0} \\
c_{0} & \cdots & c_{0} & 1
\end{pmatrix},
$$

(6)

which has the largest eigenvalue $\lambda_{0}^{(0)} = c_{0}(N-1) + 1$ and its eigenvector components $u_{i}^{(0)} = \langle \text{stock } i \rangle / \langle 0^{(0)} \rangle = 1 / \sqrt{N}$.

When a small perturbation is turned on, the total correlation matrix becomes

$$
C = C^{0} + \Delta,
$$

(7)

where $\Delta_{ij} = 0$ and $\Delta_{ij} = \Delta_{ji}$. Applying the perturbation theory up to the first order, the largest eigenvalue and the corresponding eigenvector components are easily calculated as

$$
\lambda_{0} = c_{0}(N-1) + 1 + \frac{1}{N} \sum_{ij} \Delta_{ij},
$$

$$
u_{i} = \frac{1}{c_{0}N^{3/2}} \left( w_{i} + c_{0} - \frac{1}{N} \sum_{jk} \Delta_{jk} \right),
$$

(8)

where $w_{i} = \sum_{j \neq i} C_{ij}$.

We check the validity of Eqs. (8) by comparing with the largest eigenvector obtained from the numerical diagonalization of the empirical correlation matrix. For the comparison, we make the distribution of $C_{ij}$ in Eq. (7) to be close to the empirical $C_{ij}$ distribution by assuming that $\Delta_{ij}$ follows the bell-shaped distribution with zero mean and letting $c_{0}$ to the mean value of the empirical $C_{ij}$. Because the assumption not only reproduces the distribution of empirical $C_{ij}$ but also allows us to neglect the $1/N\Sigma\Delta_{ij}$ term in Eqs. (8), we can directly compare the perturbation theory with the numerical result. Figure 4 displays the eigenvector components of the largest eigenvalue obtained from the empirical correlation matrix and the dominant terms of Eqs. (8), which show remarkable agreement with each other.

Equation (8) indicates that the eigenvector of the largest eigenvalue is contributed by not only the global fluctuation but also the unknown perturbations from $\Delta$ including random noises. Thus, by filtering out the $C^{m}$ term, we can decrease the effect of unnecessary perturbations in constructing the group correlation matrix. Indeed, as seen in Fig. 2(c), because the heavy tail part of $C^{s}$, the highly correlated elements, are buried in $C^{m}$, the clustering of stocks would be seriously disturbed unless $C^{m}$ is filtered out.

In addition, Eqs. (8) also enable us to interpret more detailed meaning of the eigenvector than the marketwide effect. Because the $i$th eigenvector component $u_{i}$ is mostly determined by $w_{i}$, the sum of the correlation over all the other stocks, it can be regarded as the influencing power of the company in the entire stock market. In real data, the top four stocks with highest $w_{i}$ are found to be General Electric (GE), American Express (AXP), Merrill Lynch (MER), and Emerson Electric (EMR), mostly conglomerates or huge financial companies, which convinces us that $u_{i}$ is indeed representing the influencing power of stock $i$. However, these high influencing companies prevent clear clustering of stocks because of their non-negligible correlations with entire stocks in the market. This is easily comprehensible by considering an analogous situation in a network where the big hub, a node with a large number of links, can make indispensable connections between groups of nodes to cause difficulties in distinguishing the groups [19]. Therefore it is very important to filter out $C^{m}$ in order to identify the groups of stocks efficiently.

III. IDENTIFICATION OF STOCK GROUPS

In the group model for stock price correlation proposed by Noh [20], the correlation matrix $C$ takes the form of $C = C^{s} + C^{r}$, where $C^{s}$ and $C^{r}$ are the correlation matrix of stock groups and random correlation matrix, respectively. The model assumes the ideal situation with $C_{ij}^{r} = \delta_{ij} \alpha_{i}^{2}$, where $\alpha_{i}$ indicates the group to which the stock $i$ belongs. Thus $C^{s}$ is the block diagonal matrix,
\[
C^p = \begin{pmatrix}
1_0 & 0 & \cdots & 0 \\
0 & 1_1 & \vdots & \\
\vdots & \ddots & \ddots & 0 \\
0 & \cdots & 0 & 1_n
\end{pmatrix},
\]

where \(1_i\) is the \(N_i \times N_i\) matrix (\(N_i\) is the number of stocks in the \(i\)th group) of which all elements are 1.

Here we use this group model to find the groups of stocks. If the correlation matrix in the real market can be represented by the block diagonal matrix as in the model, it would be very easy to identify the groups of stocks. However, there exist infinitely many possible representations of the matrix depending on indexing of rows and columns even if we have a matrix equivalent to the block diagonal matrix. For instance, if we exchange the indices of the matrix (e.g., \(\{i,j,k\} \rightarrow \{k,i,j\}\)) the matrix may not be block-diagonal anymore. Therefore the problem in identifying the groups in stock correlation matrix requires one to find out the optimized sequence of stocks to transform the matrix into the well-organized block diagonal matrix [21].

To optimize the sequence of stocks for clear block diagonalization, we consider the correlation between two stocks as an attraction force between them. For the ideal group correlation matrix in the group model, the block diagonal form is evidently the most stable form if the attractive force between stocks is proportional to their correlation within the group. Evidently the most stable form if the attractive force between stocks is proportional to their correlation within the group. For the ideal group correlation matrix, the block diagonal form is evidently the most stable form if the attractive force between stocks is proportional to their correlation within the group.

As an alternative method, we also perform a network-based approach to find the groups of stocks. In principle, the correlation matrix can be treated as an adjacency matrix of the weighted network of stocks, in which the weights indicate how closely correlated the stocks are in their price changes [26]. However, for the simplicity and the clear definition of groups in the network, we consider the binary network of stocks which permits only two possible states of a stock pair, connected or disconnected.

To construct the binary network of stocks, we use the percolation approach because of its usefulness of finding groups. The method is very simple: for each pair of stocks, we connect them if the group correlation coefficient \(C_{ij}^p\) is larger than a preassigned threshold value \(p\). If the heavy tail in the distribution of \(C_{ij}^p\) in Fig. 2 mostly comes from the correlation between the stocks in the same group, an appropriate choice of \(p=p_s\) will give several meaningful isolated clusters, \(m\), in the network which are expected to be identified as different stock groups.

We determine \(p_s\) by observing the change of the network structure as \(p\) decreases. Figure 6(a) displays the number of isolated clusters in the network as a function of threshold \(p\). As we decrease \(p\), the number of isolated clusters in the network increases slowly and stays near the maximum value up to \(p=0.1\), and then it abruptly decreases to 1, which indicates there exists only one isolated cluster. Therefore we choose \(p_s=0.1\) to construct the most clustered but stable stock network [27].

We find that the constructed network consists of separable groups of stocks which correspond to the industrial sectors of...
At $p_c = 0.1$, the network has 92 nodes and 357 links. The identification of stock group is very clear because the clusters in the network, which we consider to be equivalent to stock groups, are fully connected networks or very dense networks in which most of the nodes in the cluster are directly connected. However, although most of the stock groups are represented by a single industrial sector, it is found that the stocks which belong to two different industrial
sectors coexist in a cluster. For instance, the stocks in the healthcare sector and the noncyclical consumer sector cannot be separable in this network. Indeed, in Fig. 5, one can observe non-negligible correlation between the healthcare and the noncyclical consumer, which indicates the presence of an intergroup correlation. In the real market, this presence of such an intersector correlation can be expected and our clustering results shown in Figs. 5 and 7 present both intergroup and intragroup correlations that exist in the real stock market.

The group identification based on the eigenvector analysis of the stock price correlation matrix has been studied by several research groups \cite{9,10,11}. In spite of their pioneering achievements to reveal the localization properties of eigenvectors, the classification of stocks into groups was not so clear, and it only covered about 10\% of their stocks because they used only the few highest contributions of eigenvector components due to the ambiguity explained in Sec. I. In this work, we not only introduce a more refined and systematic method to identify the stock groups, but also successfully cluster about 70\% of stocks into groups although direct comparison of the success ratio might be inappropriate because our data set is different from theirs.

On the other hand, Onnela et al. \cite{6} introduced the percolation approach to construct the stock network in which the links are added between stocks one by one in descending order from the highest element of the full correlation matrix. In their work, though highly correlated groups of stocks were found, the threshold value of the correlation to settle the network structure was hardly determined; the number of isolated clusters according to the threshold did not show the clear cut. We believe that this is attributed to the fact that they used the full correlation matrix carrying marketwide and random fluctuation. We would also fail to determine the critical threshold value of correlation if we use the full correlation matrix instead of the filtered one [see Fig. 6(b)]. This indicates that the filtering is crucial for the stock group identification.

Finally, we note that Marsili et al. introduced a different method to filter noises from the time series of stock price log-returns for stock group identification. In their work, it was assumed that the normalized log-return could be expressed by the linear combination of the noise at individual stock level and the noise at the level of the groups, which fitted to the real data to determine the weights of two noises and the constituents of the groups. However, we found that the effect of the inhomogeneous marketwide fluctuation is quite significant that the marketwide effect needs to be considered seriously to describe the correlation between stock

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure6}
\caption{The dependence of the number of isolated clusters, $m$, in the stock network on the threshold $p$ in constructing the network from (a) the group correlation matrix and (b) the full correlation matrix.}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure7}
\caption{The stock network with the threshold $p_c=0.1$. The thickness of links indicates the strength of the correlation in the group correlation matrix.}
\end{figure}
correctly. Indeed, it is found that the filtering out of the corresponding $C_G$ improves the clustering result.

IV. CONCLUSION

In conclusion, we successfully identify the multiple group of stocks from the empirical correlation matrix of stock price changes in the New York Stock Exchange. We propose refined methods to find stock groups which dramatically reduce ambiguities as compared to identifying stock groups from the localization in a single eigenvector of the correlation matrix [9–11]. From the analysis of the characteristics of eigenvectors, we construct the group correlation matrix of the stock groups excluding the marketwide effect and random noise. By optimizing the representation of the group correlation matrix, we find that the group correlation matrix is represented by the block diagonal matrix where the stocks in each block belong to the same group. This coincides with the theoretical model of Noh [20]. Equally good stock group identification is also achieved by the percolation approach on the group correlation matrix to construct the network of stocks.

ACKNOWLEDGMENTS

We thank J. D. Noh for helpful discussions. This work was supported by Grant No. R14-2002-059-01002-0 from KOSEF-ABRL program.

[15] Our selection of stocks covers most available stock price data for the period 1983–2003 in the database. We use the adjusted daily stock prices and skip ±1 days of the data when splits and the dividends of any stock occur in Eqs. (1) and (2) to avoid the possible artifact caused by the abrupt stock price changes [see http://help.yahoo.com/help/us/fin/quote/index.html].
[17] The dynamics of the largest eigenvalue and its eigenvector has been studied very recently by using perturbation theory in the separate work [18].
[21] The similar matrix element rearranging technique was used in Ref. [22].
[23] The visualization of $C^G$ is not very sensitive to the cutoff value $c_r$.
[27] Small change of $p_r$, $\sim 0.01$, does not seriously change the clustering result.