Random matrix approach to cross correlations in financial data

Vasiliki Plerou,[1,2,*] Parameswaran Gopikrishnan,[1] Bernd Rosenow,[1,3] Luís A. Nunes Amaral,[1] Thomas Guhr,[4,5]
and H. Eugene Stanley[1]
[1]*Center for Polymer Studies and Department of Physics, Boston University, Boston, Massachusetts 02215*
[2]*Department of Physics, Boston College, Chestnut Hill, Massachusetts 02167*
[3]*Department of Physics, Harvard University, Cambridge, Massachusetts 02138*
[4]*Max-Planck-Institute for Nuclear Physics, D-69029 Heidelberg, Germany*
[5]*Matematisk Fysik, LTH, Lunds Universitet, Lund, Sweden*

(Received 1 August 2001; published 27 June 2002)

We analyze cross correlations between price fluctuations of different stocks using methods of random matrix theory (RMT). Using two large databases, we calculate cross-correlation matrices C of returns constructed from (i) 30-min returns of 1000 US stocks for the 2-yr period 1994–1995, (ii) 30-min returns of 881 US stocks for the 2-yr period 1996–1997, and (iii) 1-day returns of 422 US stocks for the 35-yr period 1962–1996. We test the statistics of the eigenvalues λ_i of C against a "null hypothesis" — a random correlation matrix constructed from mutually uncorrelated time series. We find that a majority of the eigenvalues of C fall within the RMT bounds $[\lambda_-,\lambda_+]$ for the eigenvalues of random correlation matrices. We test the eigenvalues of C within the RMT bound for universal properties of random matrices and find good agreement with the results for the Gaussian orthogonal ensemble of random matrices—implying a large degree of randomness in the measured cross-correlation coefficients. Further, we find that the distribution of eigenvector components for the eigenvectors corresponding to the eigenvalues outside the RMT bound display systematic deviations from the RMT prediction. In addition, we find that these "deviating eigenvectors" are stable in time. We analyze the components of the deviating eigenvectors and find that the largest eigenvalue corresponds to an influence common to all stocks. Our analysis of the remaining deviating eigenvectors shows distinct groups, whose identities correspond to conventionally identified business sectors. Finally, we discuss applications to the construction of portfolios of stocks that have a stable ratio of risk to return.

DOI: 10.1103/PhysRevE.65.066126 PACS number(s): 89.90.+n, 05.45.Tp, 05.40.Fb

I. INTRODUCTION

A. Motivation

Quantifying correlations between different stocks is a topic of interest not only for scientific reasons of understanding the economy as a complex dynamical system, but also for practical reasons such as asset allocation and portfolio-risk estimation [1–4]. Unlike most physical systems, where one relates correlations between subunits to basic interactions, the underlying "interactions" for the stock market problem are not known. Here, we analyze cross correlations between stocks by applying concepts and methods of random matrix theory, developed in the context of complex quantum systems where the precise nature of the interactions between subunits are not known.

In order to quantify correlations, we first calculate the price change ("return") of stock $i=1,\ldots,N$ over a time scale Δt,

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

where $S_i(t)$ denotes the price of stock *i*. Since different stocks have varying levels of volatility (standard deviation), we define a normalized return

$$g_i(t)\equiv\frac{G_i(t)-\langle G_i\rangle}{\sigma_i}, \tag{2}$$

where $\sigma_i\equiv\sqrt{\langle G_i^2\rangle-\langle G_i\rangle^2}$ is the standard deviation of G_i, and $\langle\cdots\rangle$ denotes a time average over the period studied. We then compute the equal-time cross-correlation matrix C with elements

$$C_{ij}\equiv\langle g_i(t)g_j(t)\rangle. \tag{3}$$

By construction, the elements C_{ij} are restricted to the domain $-1\leqslant C_{ij}\leqslant 1$, where $C_{ij}=1$ corresponds to perfect correlations, $C_{ij}=-1$ corresponds to perfect anticorrelations, and $C_{ij}=0$ corresponds to uncorrelated pairs of stocks.

The difficulties in analyzing the significance and meaning of the empirical cross-correlation coefficients C_{ij} are due to several reasons, which include the following:

(i) Market conditions change with time and the cross correlations that exist between any pair of stocks may not be stationary.

(ii) The finite length of time series available to estimate cross correlations introduces "measurement noise."

If we use a long-time series to circumvent the problem of finite length, our estimates will be affected by the nonstationarity of cross correlations. For these reasons, the empirically-measured cross correlations will contain "random" contributions, and it is a difficult problem in general to estimate from C the cross correlations that are not a result of randomness.

How can we identify from C_{ij}, those stocks that remained correlated (on the average) in the time period studied? To answer this question, we test the statistics of C against the "null hypothesis" of a random correlation

[*] Corresponding author. Email address: plerou@cgl.bu.edu

matrix—a correlation matrix constructed from mutually uncorrelated time series. If the properties of C conform to those of a random correlation matrix, then it follows that the contents of the empirically measured C are random. Conversely, deviations of the properties of C from those of a random correlation matrix convey information about ‘‘genuine’’ correlations. Thus, our goal shall be to compare the properties of C with those of a random correlation matrix and separate the content of C into two groups: (a) the part of C that conforms to the properties of random correlation matrices (‘‘noise’’) and (b) the part of C that deviates (‘‘information’’).

B. Background

The study of statistical properties of matrices with independent random elements—*random matrices*—has a rich history originating in nuclear physics [5–13]. In nuclear physics, the problem of interest 50 years ago was to understand the energy levels of complex nuclei, which the existing models failed to explain. Random matrix theory (RMT) was developed in this context by Wigner, Dyson, Mehta, and others in order to explain the statistics of energy levels of complex quantum systems. They postulated that the Hamiltonian describing a heavy nucleus can be described by a matrix H with independent random elements H_{ij} drawn from a probability distribution [5–9]. Based on this assumption, a series of remarkable predictions were made that are found to be in agreement with the experimental data [5–7]. For complex quantum systems, RMT predictions represent an average over all possible interactions [8–10]. Deviations from the *universal* predictions of RMT identify system specific, nonrandom properties of the system under consideration, providing clues about the underlying interactions [11–13].

Recent studies [14,15] applying RMT methods to analyze the properties of C show that $\approx 98\%$ of the eigenvalues of C agree with RMT predictions, suggesting a considerable degree of randomness in the measured cross correlations. It is also found that there are deviations from RMT predictions for $\approx 2\%$ of the largest eigenvalues. These results prompt the following questions:

(1) What is a possible interpretation for the deviations from RMT?

(2) Are the deviations from RMT stable in time?

(3) What can we infer about the structure of C from these results?

(4) What are the practical implications of these results?

In the following, we address these questions in detail. We find that the largest eigenvalue of C represents the influence of the entire market that is common to all stocks. Our analysis of the contents of the remaining eigenvalues that deviate from RMT shows the existence of cross correlations between stocks of the same type of industry, stocks having large market capitalization, and stocks of firms having business in certain geographical areas [16–18]. By calculating the scalar product of the eigenvectors from one time period to the next, we find that the ‘‘deviating eigenvectors’’ have varying degrees of time stability, quantified by the magnitude of the scalar product. The largest two to three eigenvectors are stable for extended periods of time, while for the rest of the deviating eigenvectors, the time stability decreases as the corresponding eigenvalues are closer to the RMT upper bound.

To test that the deviating eigenvalues are the only ‘‘genuine’’ information contained in C, we compare the eigenvalue statistics of C with the known universal properties of real symmetric random matrices, and we find good agreement with the RMT results. Using the notion of the inverse participation ratio, we analyze the eigenvectors of C and find large values of inverse participation ratio at both edges of the eigenvalue spectrum—suggesting a ‘‘random band’’ matrix structure for C. Last, we discuss applications to the practical goal of finding an investment that provides a given return without exposure to unnecessary risk. In addition, it is possible that our methods can also be applied for filtering out ‘‘noise’’ in empirically measured cross-correlation matrices in a wide variety of applications.

This paper is organized as follows. Section II contains a brief description of the data analyzed. Section III discusses the statistics of cross-correlation coefficients. Section IV discusses the eigenvalue distribution of C and compares with RMT results. Section V tests the eigenvalue statistics C for universal properties of real symmetric random matrices and Sec. VI contains a detailed analysis of the contents of eigenvectors that deviate from RMT. Section VII discusses the time stability of the deviating eigenvectors. Section VIII contains applications of RMT methods to construct ‘‘optimal’’ portfolios that have a stable ratio of risk to return. Finally, Sec. IX contains some concluding remarks.

II. DATA ANALYZED

We analyze two different databases covering securities from the three major US stock exchanges, namely, the New York Stock Exchange (NYSE), the American Stock Exchange (AMEX), and the National Association of Securities Dealers Automated Quotation (NASDAQ).

Database I. We analyze the Trades and Quotes (TAQ) database, that documents all transactions for all major securities listed in all the three stock exchanges. We extract from this database time series of prices [19] of the 1000 largest stocks by market capitalization on the starting date January 3, 1994. We analyze this database for the 2-yr period 1994–1995 [20]. From this database, we form $L=6448$ records of 30-min returns of $N=1000$ US stocks for the 2-yr period 1994–1995. We also analyze the prices of a subset comprising 881 stocks (of those 1000 we analyze for 1994–1995) that survived through two additional years 1996–1997. From this data, we extract $L=6448$ records of 30-min returns of $N=881$ US stocks for the 2-yr period 1996–1997.

Database II. We analyze the Center for Research in Security Prices (CRSP) database. The CRSP stock files cover common stocks listed on NYSE beginning in 1925, the AMEX beginning in 1962, and the NASDAQ beginning in 1972. The files provide complete historical descriptive information and market data including comprehensive distribution information, high, low, and closing prices, trading volumes, shares outstanding, and total returns. We analyze daily returns for the stocks that survive for the 35-yr period 1962–

1996 and extract $L=8685$ records of 1-day returns for $N=422$ stocks.

III. STATISTICS OF CORRELATION COEFFICIENTS

We analyze the distribution $P(C_{ij})$ of the elements $\{C_{ij};i\neq j\}$ of the cross-correlation matrix C. We first examine $P(C_{ij})$ for 30-min returns from the TAQ database for the 2-yr periods 1994–1995 and 1996–1997 [Fig. 1(a)]. First, we note that $P(C_{ij})$ is asymmetric and centered around a positive mean value ($\langle C_{ij}\rangle>0$), implying that positively correlated behavior is more prevalent than negatively correlated (anticorrelated) behavior. Second, we find that $\langle C_{ij}\rangle$ depends on time, e.g., the period 1996–1997 shows a larger $\langle C_{ij}\rangle$ than the period 1994–1995. We contrast $P(C_{ij})$ with a control—a correlation matrix R with elements R_{ij} constructed from $N=1000$ mutually uncorrelated time series, each of length $L=6448$, generated using the empirically found distribution of stock returns [21,22]. Figure 1(a) shows that $P(R_{ij})$ is consistent with a Gaussian with zero mean, in contrast to $P(C_{ij})$. In addition, we see that the part of $P(C_{ij})$ for $C_{ij}<0$ (which corresponds to anticorrelations) is within the Gaussian curve for the control, suggesting the possibility that the observed negative cross correlations in C may be an effect of randomness. Furthermore, our analysis of a surrogate correlation matrix generated from the randomized empirical time series of returns show good agreement with the Gaussian curve for the control [Fig. 1(b)].

Figure 1(c) shows $P(C_{ij})$ for daily returns from the CRSP database for five nonoverlapping 7-yr subperiods in the 35-yr period 1962–1996. We see that the time dependence of $\langle C_{ij}\rangle$ is more pronounced in this plot. In particular, the period containing the market crash of October 19, 1987 has the largest average value $\langle C_{ij}\rangle$, suggesting the existence of cross correlations that are more pronounced in volatile periods than in calm periods [23–25]. We test this possibility by comparing $\langle C_{ij}\rangle$ with the average volatility of the market (measured using the S&P 500 index), which shows large values of $\langle C_{ij}\rangle$ during periods of large volatility (Fig. 2).

IV. EIGENVALUE DISTRIBUTION OF THE CORRELATION MATRIX

As stated above, our aim is to extract information about cross correlations from C. So, we compare the properties of C with those of a random cross-correlation matrix [14]. In matrix notation, the correlation matrix can be expressed as

$$\mathsf{C}=\frac{1}{L}\mathsf{G}\,\mathsf{G}^T, \tag{4}$$

where G is an $N\times L$ matrix with elements $\{g_{i\,m}\equiv g_i(m\Delta t);i=1,\ldots,N;\ m=0,\ldots,L-1\}$, and G^T denotes the transpose of G. Therefore, we consider a random correlation matrix

$$\mathsf{R}=\frac{1}{L}\mathsf{A}\,\mathsf{A}^T, \tag{5}$$

where A is an $N\times L$ matrix containing N time series of L random elements $a_{i\,m}$ with zero mean and unit variance, that are mutually uncorrelated.

Statistical properties of random matrices such as R are known [26,27]. Particularly, in the limit $N\rightarrow\infty$, $L\rightarrow\infty$, such that $Q\equiv L/N$ (>1) is fixed, it was shown analytically [27] that the probability density function $P_{\rm rm}(\lambda)$ of eigenvalues λ of the random correlation matrix R is given by

$$P_{\rm rm}(\lambda)=\frac{Q}{2\pi}\frac{\sqrt{(\lambda_+-\lambda)(\lambda-\lambda_-)}}{\lambda}, \tag{6}$$

for λ within the bounds $\lambda_-\leqslant\lambda_i\leqslant\lambda_+$, where λ_- and λ_+ are the minimum and maximum eigenvalues of R, respectively, given by

$$\lambda_\pm=1+\frac{1}{Q}\pm2\sqrt{\frac{1}{Q}}. \tag{7}$$

For finite L and N, the abrupt cutoff of $P_{\rm rm}(\lambda)$ is replaced by a rapidly decaying edge [28]. Note that the expression Eq. (6) is exact for the case of Gaussian-distributed matrix elements $a_{i\,m}$. Numerically, we find that for the case of power-law distributed $a_{i\,m}$, the eigenvalue distribution of the control correlation matrix shows good agreement with Eq. (6), as long as the power-law exponents are outside the Lévy stable domain [29]. In particular, for the case of power-law distributed time series with exponent identical to that for stock returns [21,22], we find good agreement with Eq. (6).

We next compare the eigenvalue distribution $P(\lambda)$ of C with $P_{\rm rm}(\lambda)$ [14]. We examine $\Delta t=30$-min returns for $N=1000$ stocks, each containing $L=6448$ records. Thus $Q=6.448$, and we obtain $\lambda_-=0.36$ and $\lambda_+=1.94$ from Eq. (7). We compute the eigenvalues λ_i of C, where λ_i are rank ordered ($\lambda_{i+1}>\lambda_i$) [30]. Figure 3(a) compares the probability distribution $P(\lambda)$ with $P_{\rm rm}(\lambda)$ calculated for $Q=6.448$. We note the presence of a well-defined "bulk" of eigenvalues which fall within the bounds $[\lambda_-,\lambda_+]$ for $P_{\rm rm}(\lambda)$. We also note deviations for a few (≈20) largest and smallest eigenvalues. In particular, the largest eigenvalue $\lambda_{1000}\approx50$ for the 2-yr period, which is ≈25 times larger than $\lambda_+=1.94$.

Since Eq. (6) is strictly valid only for $L\rightarrow\infty$ and $N\rightarrow\infty$, we must test that the deviations that we find in Fig. 3(a) for the largest few eigenvalues are not an effect of finite values of L and N. To this end, we contrast $P(\lambda)$ with the RMT result $P_{\rm rm}(\lambda)$ for the random correlation matrix of Eq. (5), constructed from $N=1000$ mutually uncorrelated time series generated to have identical power-law tails as the empirical distribution of returns [21], each of the same length $L=6448$. We find good agreement with Eq. (6) [Fig. 3(b)], thus showing that the deviations from RMT found for the largest few eigenvalues in Fig. 3(a) are not a result of the fact that L and N are finite, or of the fact that returns are fat tailed.

As an additional test, we randomize the empirical time series of returns, thereby destroying all the equal-time correlations that exist. We then compute a surrogate correlation matrix. The eigenvalue distribution for this surrogate correlation matrix [Fig. 3(c)] shows good agreement with Eq. (6),

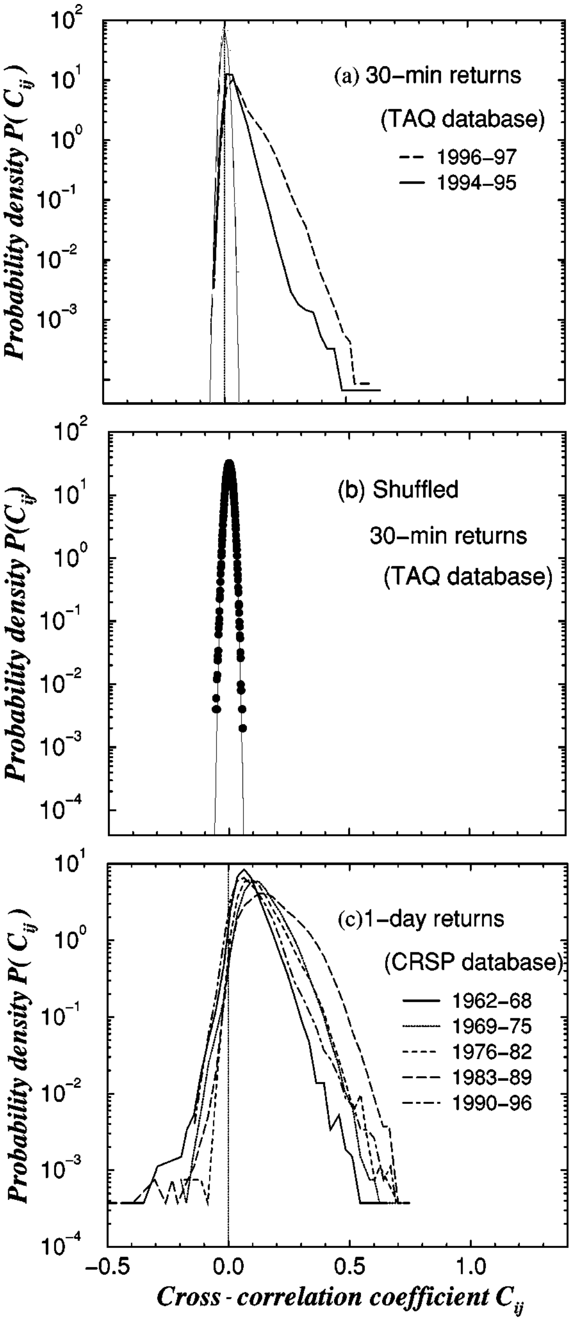

FIG. 1. (a) $P(C_{ij})$ for C calculated using 30-min returns of 1000 stocks for the 2-yr period 1994–1995 (solid line) and 881 stocks for the 2-yr period 1996–1997 (dashed line). For the period 1996–1997 $\langle C_{ij}\rangle = 0.06$, larger than the value $\langle C_{ij}\rangle = 0.03$ for 1994–1995. The narrow parabolic curve shows the distribution of correlation coefficients for the control $P(R_{ij})$ of Eq. (5), which is consistent with a Gaussian distribution with zero mean. (b) $P(C_{ij})$ (circles) for the correlation matrix calculated using randomized 30-min returns of 1000 stocks (1994-1995) shows good agreement with the control (solid curve). (c) $P(C_{ij})$ calculated from daily returns of 422 stocks for five 7-yr subperiods in the 35 years 1962–1996. We find a large value of $\langle C_{ij}\rangle = 0.18$ for the period 1983–1989, compared with the average $\langle C_{ij}\rangle = 0.10$ for the other periods.

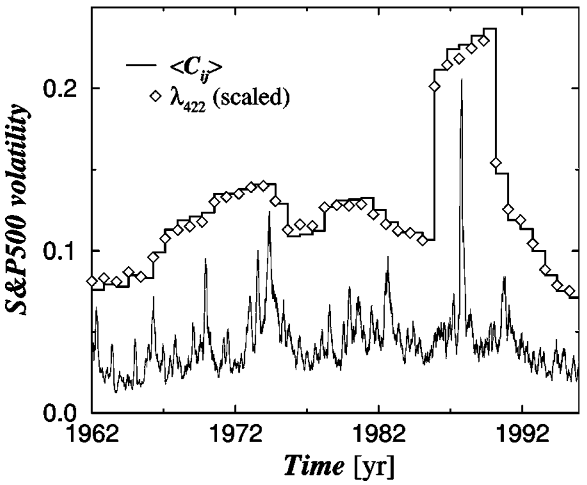

FIG. 2. The stair-step curve shows the average value of the correlation coefficients $\langle C_{ij}\rangle$, calculated from 422×422 correlation matrices C constructed from daily returns using a sliding $L=965$ day time window in discrete steps of $L/5=193$ days. The diamonds correspond to the largest eigenvalue λ_{422} (scaled by a factor 4×10^2) for the correlation matrices thus obtained. The bottom curve shows the S&P 500 volatility (scaled for clarity) calculated from daily records with a sliding window of length 40 days. We find that both $\langle C_{ij}\rangle$ and λ_{422} have large values for periods containing the market crash of October 19, 1987.

confirming that the largest eigenvalues in Fig. 3(a) are genuine effect of equal-time correlations among stocks.

Figure 4 compares $P(\lambda)$ for C calculated using $L=1737$ daily returns of 422 stocks for the 7-yr period 1990–1996. We find a well-defined bulk of eigenvalues that fall within $P_{\rm rm}(\lambda)$, and deviations from $P_{\rm rm}(\lambda)$ for large eigenvalues—similar to what we found for $\Delta t=30$ min [Fig. 3(a)]. Thus, a comparison of $P(\lambda)$ with the RMT result $P_{\rm rm}(\lambda)$ allows us to distinguish the *bulk* of the eigenvalue spectrum of C that agrees with RMT (random correlations) from the deviations (genuine correlations).

V. UNIVERSAL PROPERTIES: ARE THE BULK OF EIGENVALUES OF C CONSISTENT WITH RMT?

The presence of a well-defined bulk of eigenvalues that agree with $P_{\rm rm}(\lambda)$ suggests that the contents of C are mostly random except for the eigenvalues that deviate. Our conclusion was based on the comparison of the eigenvalue distribution $P(\lambda)$ of C with that of random matrices of the type $\mathsf{R}=(1/L)\mathsf{A}\mathsf{A}^T$. Quite generally, comparison of the eigenvalue distribution with $P_{\rm rm}(\lambda)$ alone is not sufficient to support the possibility that the bulk of the eigenvalue spectrum of C is random. Random matrices that have drastically different $P(\lambda)$ share similar correlation structures in their eigenvalues—universal properties—that depend only on the general symmetries of the matrix [11–13]. Conversely, matrices that have the same eigenvalue distribution can have drastically different eigenvalue correlations. Therefore, a test of randomness of C involves the investigation of correlations in the eigenvalues λ_i.

Since by definition C is a real symmetric matrix, we shall

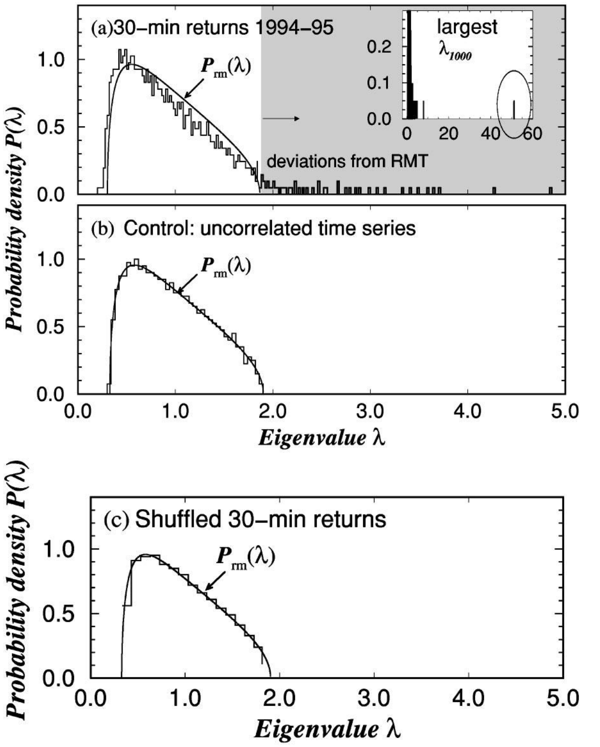

FIG. 3. (a) Eigenvalue distribution $P(\lambda)$ for C constructed from the 30-min returns for 1000 stocks for the 2-yr period 1994–1995. The solid curve shows the RMT result $P_{\rm rm}(\lambda)$ of Eq. (6). We note several eigenvalues outside the RMT upper bound λ_+ (shaded region). The inset shows the largest eigenvalue $\lambda_{1000}\approx 50 \gg \lambda_+$. (b) $P(\lambda)$ for the random correlation matrix R, computed from $N=1000$ computer-generated random uncorrelated time series with length $L=6448$ shows good agreement with the RMT result, Eq. (6) (solid curve). (c) Eigenvalue distribution for a surrogate correlation matrix constructed from randomized 30-min returns shows good agreement with Eq. (6) (solid curve).

test the eigenvalue statistics C for universal features of eigenvalue correlations displayed by real symmetric random matrices. Consider a $M\times M$ real symmetric random matrix S with off-diagonal elements S_{ij}, which for $i<j$ are independent and identically distributed with zero mean $\langle S_{ij}\rangle = 0$ and variance $\langle S_{ij}^2\rangle > 0$. It is conjectured based on analytical [31] and extensive numerical evidence [11] that in the limit $M\to\infty$, regardless of the distribution of elements S_{ij}, this class of matrices, on the scale of local mean eigenvalue spacing, display the universal properties (eigenvalue correlation functions) of the ensemble of matrices whose elements are distributed according to a Gaussian probability measure—called the Gaussian orthogonal ensemble (GOE) [11].

Formally, GOE is defined on the space of real symmetric matrices by two requirements [11]. The first is that the ensemble is invariant under orthogonal transformations, i.e., for any GOE matrix Z, the transformation $\mathsf{Z}\to Z' \equiv \mathsf{W}^T\mathsf{Z}\,\mathsf{W}$, where W is any real orthogonal matrix ($\mathsf{W}\mathsf{W}^T=\mathsf{I}$), leaves the joint probability $P(Z)dZ$ of elements Z_{ij} unchanged: $P(Z')dZ' = P(Z)dZ$. The second requirement is that the elements $\{Z_{ij}; i\leqslant j\}$ are statistically independent [11].

By definition, random cross-correlation matrices R [Eq.

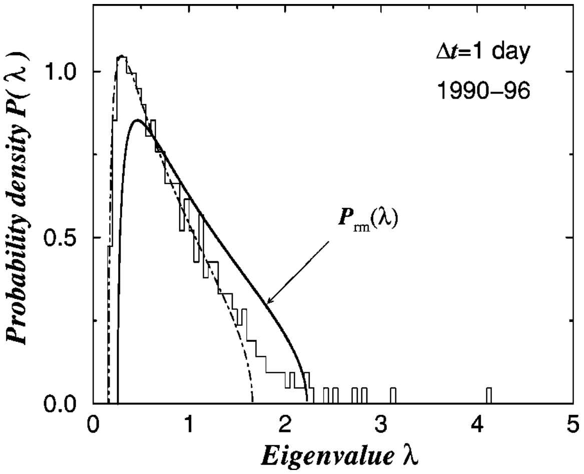

FIG. 4. $P(\lambda)$ for C constructed from daily returns of 422 stocks for the 7-yr period 1990–1996. The solid curve shows the RMT result $P_{\rm rm}(\lambda)$ of Eq. (6]) using $N=422$ and $L=1737$. The dot-dashed curve shows a fit to $P(\lambda)$ using $P_{\rm rm}(\lambda)$ with λ_+ and λ_- as free parameters. We find similar results as found in Fig. 3(a) for 30-min returns. The largest eigenvalue (not shown) has the value $\lambda_{422}=46.3$.

(5)] that we are interested in are not strictly GOE-type matrices, but rather belong to a special ensemble called the "chiral" GOE [13,32]. This can be seen by the following argument. Define a matrix B,

$$\mathsf{B} \equiv \begin{bmatrix} 0 & \mathsf{A}/\sqrt{\mathsf{L}} \\ \mathsf{A}^T/\sqrt{\mathsf{L}} & 0 \end{bmatrix}. \tag{8}$$

The eigenvalues γ of B are given by $\det(\gamma^2\mathsf{I} - \mathsf{A}\mathsf{A}^T/\mathsf{L}) = 0$ and similarly, the eigenvalues λ of R are given by $\det(\lambda\mathsf{I} - \mathsf{A}\mathsf{A}^T/\mathsf{L}) = 0$. Thus, all nonzero eigenvalues of B occur in pairs, i.e., for every eigenvalue λ of R, $\gamma_\pm = \pm\sqrt{\lambda}$ are eigenvalues of B. Since the eigenvalues occur pairwise, the eigenvalue spectra of both B and R have special properties in the neighborhood of zero that are different from the standard GOE [13,32]. As these special properties decay rapidly as one goes further from zero, the eigenvalue correlations of R in the bulk of the spectrum are still consistent with those of the standard GOE. Therefore, our goal shall be to test the bulk of the eigenvalue spectrum of the empirically measured cross-correlation matrix C with the known universal features of standard GOE-type matrices.

In the following, we test the statistical properties of the eigenvalues of C for three known universal properties [11–13] displayed by GOE matrices: (i) the distribution of nearest-neighbor eigenvalue spacings $P_{\rm nn}(s)$, (ii) the distribution of next-nearest-neighbor eigenvalue spacings $P_{\rm nnn}(s)$, and (iii) the "number variance" statistic Σ^2.

The analytical results for the three properties listed above hold if the spacings between adjacent eigenvalues (rank ordered) are expressed in units of *average* eigenvalue spacing. Quite generally, the average eigenvalue spacing changes from one part of the eigenvalue spectrum to the next. So, in order to ensure that the eigenvalue spacing has a uniform *average* value throughout the spectrum, we must find a transformation called “unfolding,” which maps the eigenvalues λ_i to new variables called “unfolded eigenvalues” ξ_i, whose distribution is uniform [11–13]. Unfolding ensures that the distances between eigenvalues are expressed in units of local mean eigenvalue spacing [11], and thus facilitates comparison with theoretical results. The procedures that we use for unfolding the eigenvalue spectrum are discussed in the Appendix.

A. Distribution of nearest-neighbor eigenvalue spacings

We first consider the eigenvalue spacing distribution, which reflects two point as well as eigenvalue correlation functions of all orders. We compare the eigenvalue spacing distribution of C with that of GOE random matrices. For GOE matrices, the distribution of “nearest-neighbor” eigenvalue spacings $s \equiv \xi_{k+1} - \xi_k$ is given by [11–13]

$$P_{\rm GOE}(s) = \frac{\pi s}{2} \exp\left(-\frac{\pi}{4} s^2 \right), \qquad (9)$$

often referred to as the “Wigner surmise” [33]. The Gaussian decay of $P_{\rm GOE}(s)$ for large s [bold curve in Fig. 5(a)] implies that $P_{\rm GOE}(s)$ “probes” scales only of the order of one eigenvalue spacing. Thus, the spacing distribution is known to be robust across different unfolding procedures [13].

We first calculate the distribution of the “nearest-neighbor spacings” $s \equiv \xi_{k+1} - \xi_k$ of the unfolded eigenvalues obtained using the Gaussian broadening procedure. Figure 5(a) shows that the distribution $P_{\rm nn}(s)$ of nearest-neighbor eigenvalue spacings for C constructed from 30-min returns for the 2-yr period 1994–1995 agrees well with the RMT result $P_{\rm GOE}(s)$ for GOE matrices.

Identical results are obtained when we use the alternative unfolding procedure of fitting the eigenvalue distribution. In addition, we test the agreement of $P_{\rm nn}(s)$ with RMT results by fitting $P_{\rm nn}(s)$ to the one-parameter Brody distribution [12,13]

$$P_{\rm Br}(s) = B(1+\beta) s^{\beta} \exp(-B s^{1+\beta}), \qquad (10)$$

where $B \equiv \{\Gamma([\beta+2]/[\beta+1])\}^{1+\beta}$. The case $\beta=1$ corresponds to the GOE and $\beta=0$ corresponds to uncorrelated eigenvalues (Poisson-distributed spacings). We obtain $\beta = 0.99 \pm 0.02$, in good agreement with the GOE prediction $\beta=1$. To test nonparametrically that $P_{\rm GOE}(s)$ is the correct description for $P_{\rm nn}(s)$, we perform the Kolmogorov-Smirnov test. We find that at the 80% confidence level, a Kolmogorov-Smirnov test cannot reject the hypothesis that the GOE is the correct description for $P_{\rm nn}(s)$.

Next, we analyze the nearest-neighbor spacing distribution $P_{\rm nn}(s)$ for C constructed from daily returns for four 7-yr periods (Fig. 6). We find good agreement with the GOE result of Eq. (9), similar to what we find for C constructed from 30-min returns. We also test that both of the unfolding procedures discussed in the Appendix yield consistent results. Thus, we have seen that the eigenvalue-spacing distribution of empirically measured cross-correlation matrices C is consistent with the RMT result for real symmetric random matrices.

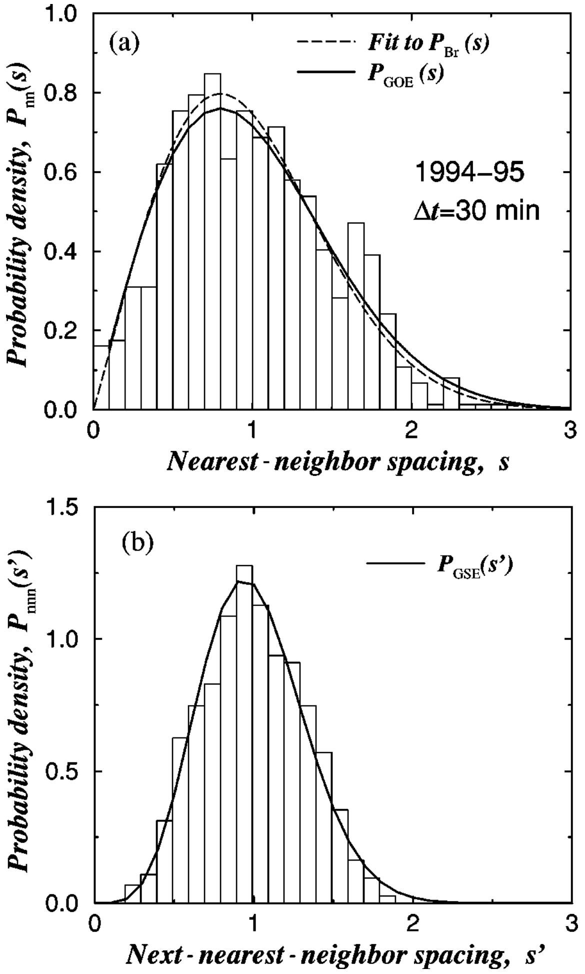

FIG. 5. (a) Nearest-neighbor (nn) spacing distribution $P_{\rm nn}(s)$ of the unfolded eigenvalues ξ_i of C constructed from 30-min returns for the 2-yr period 1994–1995. We find good agreement with the GOE result $P_{\rm GOE}(s)$ [Eq. (9)] (solid line). The dashed line is a fit to the one-parameter Brody distribution $P_{\rm Br}$ [Eq. (10)]. The fit yields $\beta = 0.99 \pm 0.02$, in good agreement with the GOE prediction $\beta=1$. A Kolmogorov-Smirnov test shows that the GOE is 10^5 times more likely to be the correct description than the Gaussian unitary ensemble, and 10^{20} times more likely than the GSE. (b) Next-nearest-neighbor (nnn) eigenvalue-spacing distribution $P_{\rm nnn}(s)$ of C compared to the nearest-neighbor spacing distribution of GSE shows good agreement. A Kolmogorov-Smirnov test cannot reject the hypothesis that $P_{\rm GSE}(s)$ is the correct distribution at the 40% confidence level. The results shown above are using the Gaussian broadening procedure. Using the second procedure of fitting $F(\lambda)$ (appendix) yields similar results.

B. Distribution of next-nearest-neighbor eigenvalue spacings

A second independent test for GOE is the distribution $P_{\rm nnn}(s')$ of *next*-nearest-neighbor spacings $s' \equiv \xi_{k+2} - \xi_k$ between the unfolded eigenvalues. For matrices of the GOE type, according to a theorem due to Ref. [10], the next-nearest-neighbor spacings follow the statistics of the Gaussian symplectic ensemble (GSE) [11–13,34]. In particular, the distribution of next-nearest-neighbor spacings $P_{\rm nnn}(s')$ for a GOE matrix is identical to the distribution of nearest-neighbor spacings of the Gaussian symplectic ensemble

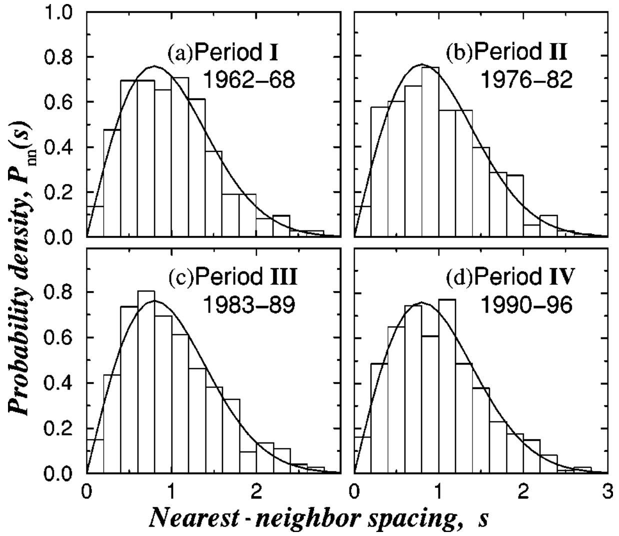

FIG. 6. Nearest-neighbor spacing distribution $P(s)$ of the unfolded eigenvalues ξ_i of C computed from the daily returns of 422 stocks for the 7-yr periods (a) 1962–1968, (b) 1976–1982, (c) 1983–1989, and (d) 1990–1996. We find good agreement with the GOE result (solid curve). The unfolding was performed by using the procedure of fitting the cumulative distribution of eigenvalues (appendix). Gaussian broadening procedure also yields similar results.

(GSE) [11,13]. Figure 5(b) shows that $P_{\mathrm{nnn}}(s')$ for the same data as Fig. 5(a) agrees well with the RMT result for the distribution of nearest-neighbor spacings of GSE matrices,

$$P_{\mathrm{GSE}}(s)=\frac{2^{18}}{3^6\pi^3}s^4\exp\left(-\frac{64}{9\pi}s^2\right). \tag{11}$$

We find that at the 40% confidence level, a Kolmogorov-Smirnov test cannot reject the hypothesis that the GSE is the correct description for $P_{\mathrm{nnn}}(s)$.

C. Long-range eigenvalue correlations

To probe for larger scales, pair correlations ("two-point" correlations) in the eigenvalues, we use the statistic Σ^2 often called the "number variance," which is defined as the variance of the number of unfolded eigenvalues in intervals of length ℓaround each ξ_i [11–13],

$$\Sigma^2(\ell)\equiv\langle[n(\xi,\ell)-\ell]^2\rangle_\xi, \tag{12}$$

where $n(\xi,\ell)$ is the number of unfolded eigenvalues in the interval $[\xi-\ell/2,\ \xi+\ell/2]$ and $\langle\cdots\rangle_\xi$ denotes an average over all ξ. If the eigenvalues are uncorrelated, $\Sigma^2\sim\ell$. For the opposite extreme of a "rigid" eigenvalue spectrum (e.g., simple harmonic oscillator), Σ^2 is a constant. Quite generally, the number variance Σ^2 can be expressed as

$$\Sigma^2(\ell)=\ell-2\int_0^\ell(\ell-x)Y(x)dx, \tag{13}$$

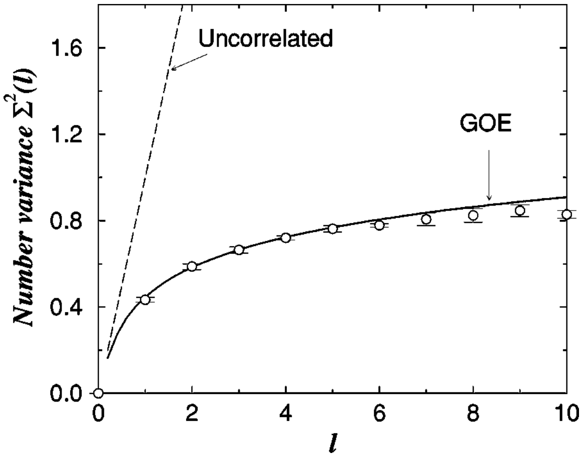

FIG. 7. (a) Number variance $\Sigma^2(\ell)$ calculated from the unfolded eigenvalues ξ_i of C constructed from 30-min returns for the 2-yr period 1994–1995. We used Gaussian broadening procedure with the broadening parameter $a=15$. We find good agreement with the GOE result of Eq. (13) (solid curve). The dashed line corresponds to the uncorrelated case (Poisson). For the range of ℓ shown, unfolding by fitting also yields similar results.

where $Y(x)$ (called "two-level cluster function") is related to the two-point correlation function [c.f., Ref. [11], p.79]. For the GOE case, $Y(x)$ is explicitly given by

$$Y(x)\equiv s^2(x)+\frac{ds}{dx}\int_x^\infty s(x')dx', \tag{14}$$

where

$$s(x)\equiv\frac{\sin(\pi x)}{\pi x}. \tag{15}$$

For large values of ℓ, the number variance Σ^2 for GOE has the "intermediate" behavior

$$\Sigma^2\sim\ln\ell. \tag{16}$$

Figure 7 shows that $\Sigma^2(\ell)$ for C calculated using 30-min returns for 1994–1995 agrees well with the RMT result of Eq. (13). For the range of ℓ shown in Fig. 7, both unfolding procedures yield similar results. Consistent results are obtained for C constructed from daily returns.

D. Implications

To summarize this section, we have tested the statistics of C for universal features of eigenvalue correlations displayed by GOE matrices. We have seen that the distribution of the nearest-neighbor spacings $P_{\mathrm{nn}}(s)$ is in good agreement with the GOE result. To test whether the eigenvalues of C display the RMT results for long-range two-point eigenvalue correlations, we analyzed the number variance Σ^2 and found good agreement with GOE results. Moreover, we also find that the statistics of next-nearest-neighbor spacings conform to the predictions of RMT. These findings show that the statistics of the *bulk* of the eigenvalues of the empirical cross-correlation matrix C is consistent with those of a real symmetric random matrix. Thus, information about genuine correlations are

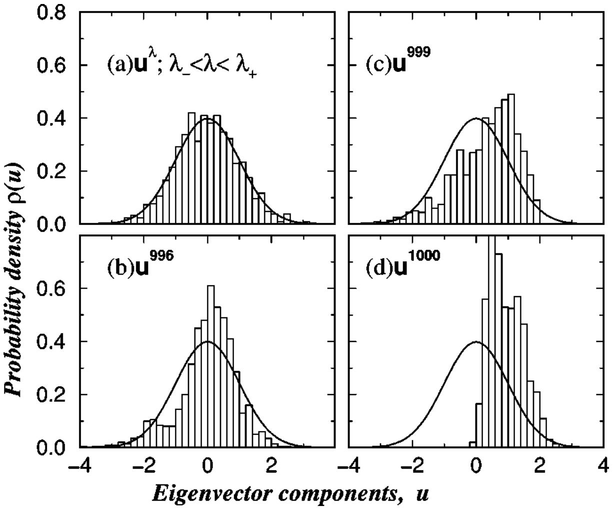

FIG. 8. (a) Distribution $\rho(u)$ of eigenvector components for one eigenvalue in the bulk $\lambda_{-}<\lambda<\lambda_{+}$ shows good agreement with the RMT prediction of Eq. (17) (solid curve). Similar results are obtained for other eigenvalues in the bulk. $\rho(u)$ for (b) u^{996} and (c) u^{999}, corresponding to eigenvalues larger than the RMT upper bound λ_{+} (shaded region in Fig. 3). (d) $\rho(u)$ for u^{1000} deviates significantly from the Gaussian prediction of RMT. The above plots are for C constructed from 30-min returns for the 2-yr period 1994–1995. We also obtain similar results for C constructed from daily returns.

contained in the deviations from RMT, which we analyze below.

VI. STATISTICS OF EIGENVECTORS

A. Distribution of eigenvector components

The deviations of $P(\lambda)$ from the RMT result $P_{\mathrm{rm}}(\lambda)$ suggests that these deviations should also be displayed in the statistics of the corresponding eigenvector components [14]. Accordingly, in this section, we analyze the distribution of eigenvector components. The distribution of the components $\{u_l^k; l=1,\ldots,N\}$ of eigenvector u^k of a random correlation matrix R should conform to a Gaussian distribution with mean zero and unit variance [13],

$$\rho_{\mathrm{rm}}(u)=\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-u^2}{2}\right). \tag{17}$$

First, we compare the distribution of eigenvector components of C with Eq. (17). We analyze $\rho(u)$ for C computed using 30-min returns for 1994–1995. We choose one typical eigenvalue λ_k from the bulk ($\lambda_{-}\leqslant\lambda_k\leqslant\lambda_{+}$) defined by $P_{\mathrm{rm}}(\lambda)$ of Eq. (6). Figure 8(a) shows that $\rho(u)$ for a typical u^k from the bulk shows good agreement with the RMT result $\rho_{\mathrm{rm}}(u)$. Similar analysis on the other eigenvectors belonging to eigenvalues within the bulk yields consistent results, in agreement with the results of the previous sections that the bulk agrees with random matrix predictions. We test the agreement of the distribution $\rho(u)$ with $\rho_{\mathrm{rm}}(u)$ by calculating the kurtosis, which for a Gaussian has the value 3. We find significant deviations from $\rho_{\mathrm{rm}}(u)$ for ≈ 20 largest and smallest eigenvalues. The remaining eigenvectors have values of kurtosis that are consistent with the Gaussian value 3.

Consider next the "deviating" eigenvalues λ_i, larger than the RMT upper bound, $\lambda_i>\lambda_{+}$. Figures 8(b) and 8(c) show that, for deviating eigenvalues, the distribution of eigenvector components $\rho(\mathsf{u})$ deviates systematically from the RMT result $\rho_{\mathrm{rm}}(\mathsf{u})$. Finally, we examine the distribution of the components of the eigenvector u^{1000} corresponding to the largest eigenvalue λ_{1000}. Figure 8(d) shows that $\rho(\mathsf{u}^{1000})$ deviates remarkably from a Gaussian, and is approximately uniform, suggesting that all stocks participate. In addition, we find that almost all components of u^{1000} have the same sign, thus causing $\rho(\mathsf{u})$ to shift to one side. This suggests that the significant participants of eigenvector u^k have a common component that affects all of them with the same bias.

B. Interpretation of the largest eigenvalue and the corresponding eigenvector

Since all components participate in the eigenvector corresponding to the largest eigenvalue, it represents an influence that is common to all stocks. Thus, the largest eigenvector quantifies the qualitative notion that certain newsbreaks (e.g., an interest rate increase) affect all stocks alike [4]. One can also interpret the largest eigenvalue and its corresponding eigenvector as the collective "response" of the entire market to stimuli. We quantitatively investigate this notion by comparing the projection (scalar product) of the time series G on the eigenvector u^{1000}, with a standard measure of US stock market performance—the returns $G_{\mathrm{SP}}(t)$ of the S&P 500 index. We calculate the projection $G^{1000}(t)$ of the time series $G_j(t)$ on the eigenvector u^{1000},

$$G^{1000}(t)\equiv\sum_{j=1}^{1000}u_j^{1000}G_j(t). \tag{18}$$

By definition, $G^{1000}(t)$ shows the return of the portfolio defined by u^{1000}. We compare $G^{1000}(t)$ with $G_{\mathrm{SP}}(t)$, and find remarkably similar behavior for the two, indicated by a large value of the correlation coefficient $\langle G_{\mathrm{SP}}(t)G^{1000}(t)\rangle=0.85$. Figure 9 shows $G^{1000}(t)$ regressed against $G_{SP}(t)$, which shows relatively narrow scatter around a linear fit. Thus, we interpret the eigenvector u^{1000} as quantifying market-wide influences on all stocks [14,15].

We analyze C at larger time scales of $\Delta t=1$ day and find similar results as above, suggesting that similar correlation structures exist for quite different time scales. Our results for the distribution of eigenvector components agree with those reported in Ref. [14], where $\Delta t=1$-day returns are analyzed. We next investigate how the largest eigenvalue changes as a function of time. Figure 2 shows the time dependence [35] of the largest eigenvalue (λ_{422}) for the 35-yr period 1962–1996. We find large values of the largest eigenvalue during periods of high market volatility, which suggests strong collective behavior in regimes of high volatility.

One way of statistically modeling an influence that is common to all stocks is to express the return G_i of stock i as

$$G_i(t)=\alpha_i+\beta_i M(t)+\epsilon_i(t), \tag{19}$$

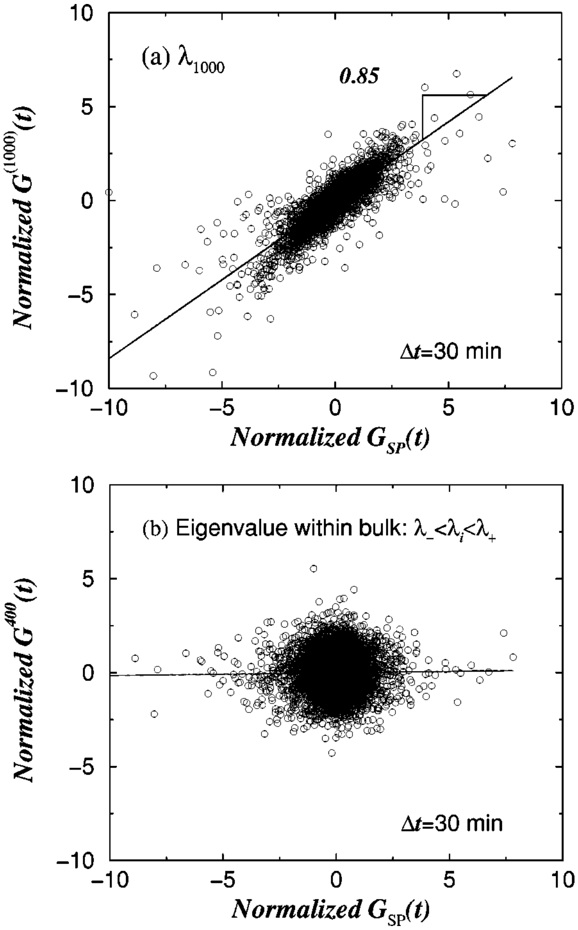

FIG. 9. (a) S&P 500 returns at $\Delta t = 30$ min regressed against the 30-min return on the portfolio G^{1000} [Eq. (18)] defined by the eigenvector u^{1000}, for the 2-yr period 1994–1995. Both axes are scaled by their respective standard deviations. A linear regression yields a slope 0.85 ± 0.09. (b) Return (in units of standard deviations) on the portfolio defined by an eigenvector corresponding to an eigenvalue λ_{400} within the RMT bounds regressed against the normalized returns of the S&P 500 index shows no significant dependence. Both axes are scaled by their respective standard deviations. The slope of the linear fit is 0.014 ± 0.011, close to 0.

where $M(t)$ is an additive term that is the same for all stocks, $\langle \epsilon(t) \rangle = 0$, α_i and β_i are stock-specific constants, and $\langle M(t)\epsilon(t) \rangle = 0$. This common term $M(t)$ gives rise to correlations between any pair of stocks. The decomposition of Eq. (19) forms the basis of widely used economic models, such as multifactor models and the Capital Asset Pricing model [4,36–52]. Since u^{1000} represents an influence that is common to all stocks, we can approximate the term $M(t)$ with $G^{1000}(t)$. The parameters α_i and β_i can therefore be estimated by an ordinary least squares regression.

Next, we remove the contribution of $G^{1000}(t)$ to each time series $G_i(t)$, and construct C from the residuals $\epsilon_i(t)$ of Eq. (19). Figure 10 shows that the distribution $P(C_{ij})$ thus obtained has significantly smaller average value $\langle C_{ij} \rangle$, showing that a large degree of cross correlations contained in C can be attributed to the influence of the largest eigenvalue (and its corresponding eigenvector) [53,54].

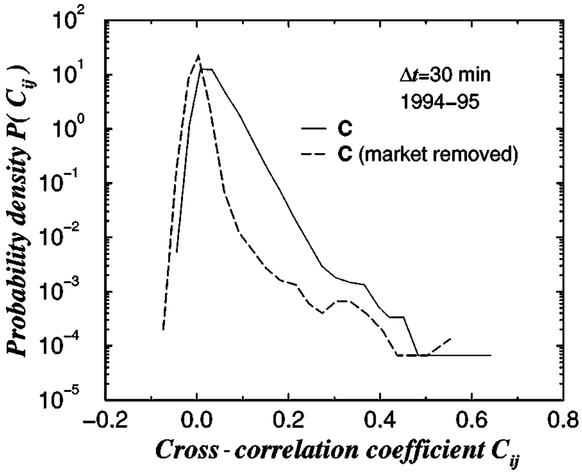

FIG. 10. Probability distribution $P(C_{ij})$ of the cross-correlation coefficients for the 2-yr period 1994–1995 before and after removing the effect of the largest eigenvalue λ_{1000}. Note that removing the effect of λ_{1000} shifts $P(C_{ij})$ toward a smaller average value $\langle C_{ij} \rangle = 0.002$ compared to the original value $\langle C_{ij} \rangle = 0.03$.

C. Number of significant participants in an eigenvector: Inverse participation ratio

Having studied the interpretation of the largest eigenvalue that deviates significantly from RMT results, we next focus on the remaining eigenvalues. The deviations of the distribution of components of an eigenvector u^k from the RMT prediction of a Gaussian is more pronounced as the separation from the RMT upper bound $\lambda_k - \lambda_+$ increases. Since proximity to λ_+ increases the effects of randomness, we quantify the number of components that participate significantly in each eigenvector, which in turn reflects the degree of deviation from RMT result for the distribution of eigenvector components. To this end, we use the notion of the inverse participation ratio (IPR), often applied in localization theory [13,55]. The IPR of the eigenvector u^k is defined as

$$I^k \equiv \sum_{l=1}^{N} [u_l^k]^4, \tag{20}$$

where u_l^k, $l = 1, \ldots, 1000$ are the components of eigenvector $\mathbf{u}^k$. The meaning of I^k can be illustrated by two limiting cases: (i) a vector with identical components $u_l^k \equiv 1/\sqrt{N}$ has $I^k = 1/N$, whereas (ii) a vector with one component $u_1^k = 1$ and the remainder zero has $I^k = 1$. Thus, the IPR quantifies the reciprocal of the number of eigenvector components that contribute significantly.

Figure 11(a) shows I^k for the case of the control of Eq. (5) using time series with the empirically found distribution of returns [21]. The average value of I^k is $\langle I \rangle \approx 3 \times 10^{-3} \approx 1/N$ with a narrow spread, indicating that the vectors are *extended* [55,56]—i.e., almost all components contribute to them. Fluctuations around this average value are confined to a narrow range (standard deviation of 1.5×10^{-4}).

Figure 11(b) shows that I^k for C constructed from 30-min returns from the period 1994–1995, agrees with I^k of the random control in the bulk ($\lambda_- < \lambda_i < \lambda_+$). In contrast, the

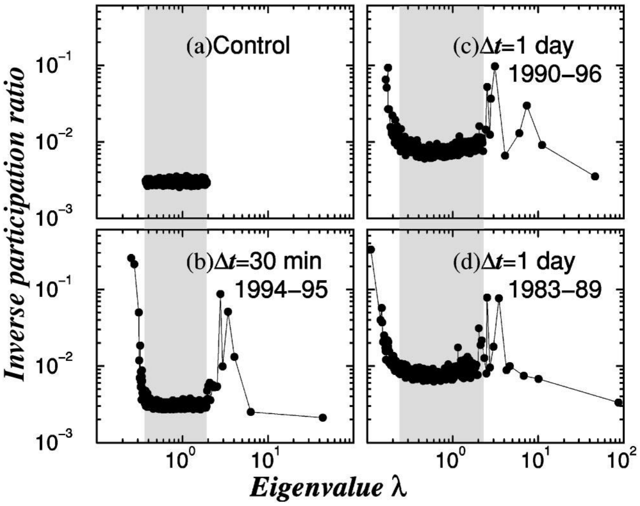

FIG. 11. (a) Inverse participation ratio (IPR) as a function of eigenvalue λ for the random cross-correlation matrix R of Eq. (6) constructed using $N=1000$ mutually uncorrelated time series of length $L=6448$. IPR for C constructed from (b) 6448 records of 30-min returns for 1000 stocks for the 2-yr period 1994–1995, (c) 1737 records of 1-day returns for 422 stocks in the 7-yr period 1990–1996, and (d) 1737 records of 1-day returns for 422 stocks in the 7-yr period 1983–1989. The shaded regions show the RMT bounds $[\lambda_+,\lambda_-]$.

edges of the eigenvalue spectrum of C show significant deviations of I^k from $\langle I\rangle$. The largest eigenvalue has $1/I^k \approx 600$ for the 30-min data [Fig. 11(b)] and $1/I^k \approx 320$ for the 1-day data [Figs. 11(c) and 11(d)], showing that almost all stocks participate in the largest eigenvector. For the rest of the large eigenvalues which deviate from the RMT upper bound, I^k values are approximately four to five times larger than $\langle I\rangle$, showing that there are varying numbers of stocks contributing to these eigenvectors. In addition, we also find that there are large I^k values for vectors corresponding to few of the small eigenvalues $\lambda_i \approx 0.25 < \lambda_-$. The deviations at both edges of the eigenvalue spectrum are considerably larger than $\langle I\rangle$, which suggests that the vectors are *localized* [55,56]—i.e., only a few stocks contribute to them.

The presence of vectors with large values of I^k also arises in the theory of Anderson localization [57]. In the context of localization theory, one frequently finds “random band matrices” [55] containing extended states with small I^k in the bulk of the eigenvalue spectrum, whereas edge states are localized and have large I^k. Our finding of localized states for small and large eigenvalues of the cross-correlation matrix C is reminiscent of Anderson localization and suggests that C may have a random band matrix structure. A random band matrix B has elements B_{ij} independently drawn from different probability distributions. These distributions are often taken to be Gaussian parametrized by their variance, which depends on i and j. Although such matrices are random, they still contain probabilistic information arising from the fact that a metric can be defined on their set of indices i. A related, but distinct way of analyzing cross correlations by defining “ultrametric” distances has been studied in Ref. [16].

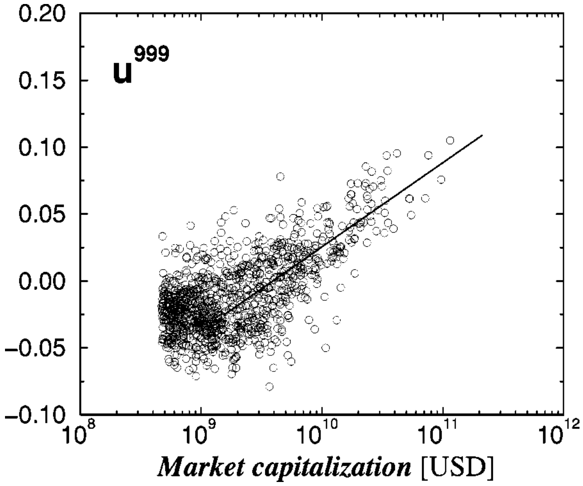

FIG. 12. All 10^3 eigenvector components of u^{999} plotted against market capitalization (in units of U.S. dollars) shows that firms with large market capitalization contribute significantly. The straight line, which shows a logarithmic fit, is a guide to the eye.

D. Interpretation of deviating eigenvectors $\mathsf{u}^{990}-\mathsf{u}^{999}$

We quantify the number of significant participants of an eigenvector using the IPR, and we examine the $1/I^k$ components of eigenvector u^k for common features [17]. A direct examination of these eigenvectors, however, does not yield a straightforward interpretation of their economic relevance. To interpret their meaning, we note that the largest eigenvalue is an order of magnitude larger than the others, which constrains the remaining $N-1$ eigenvalues since $\mathrm{Tr}\,\mathsf{C}=N$. Thus, in order to analyze the deviating eigenvectors, we must remove the effect of the largest eigenvalue λ_{1000}.

In order to avoid the effect of λ_{1000}, and thus $G^{1000}(t)$, on the returns of each stock $G_i(t)$, we perform the regression of Eq. (19), and compute the residuals $\epsilon_i(t)$. We then calculate the correlation matrix C using $\epsilon_i(t)$ in Eq. (2) and Eq. (3). Next, we compute the eigenvectors u^k of C thus obtained, and analyze their significant participants. The eigenvector u^{999} contains approximately $1/I^{999}=300$ significant participants, which are all stocks with large values of market capitalization. Figure 12 shows that the magnitude of the eigenvector components of u^{999} shows an approximately logarithmic dependence on the market capitalizations of the corresponding stocks.

We next analyze the significant contributors of the rest of the eigenvectors. We find that each of these deviating eigenvectors contains stocks belonging to similar or related industries as significant contributors. Table I shows the ticker symbols and industry groups [Standard Industry Classification (SIC) code] for stocks corresponding to the ten largest eigenvector components of each eigenvector. We find that these eigenvectors partition the set of all stocks into distinct groups that contain stocks with large market capitalization (u^{999}), stocks of firms in the electronics and computer industry (u^{998}), a combination of gold mining and investment firms (u^{996} and u^{997}), banking firms (u^{994}), oil and gas refining and equipment (u^{993}), auto manufacturing firms (u^{992}), drug manufacturing firms (u^{991}), and paper manufacturing (u^{990}).

TABLE I. Largest ten components of the eigenvectors u^{999} up to u^{991}. The columns show ticker symbols, industry type, and the standard industry classification (SIC) code, respectively.

Ticker	Industry	Industry code
	u^{999}	
XON	Oil & gas equipment/services	2911
PG	Cleaning products	2840
JNJ	Drug manufacturers/major	2834
KO	Beverages-soft drinks	2080
PFE	Drug manufacturers/major	2834
BEL	Telecom services/domestic	4813
MOB	Oil & gas equipment/services	2911
BEN	Asset management	6282
UN	Food—major diversified	2000
AIG	Property/casualty insurance	6331
	u^{998}	
TXN	Semiconductor—broad line	3674
MU	Semiconductor—memory chips	3674
LSI	Semiconductor—specialized	3674
MOT	Electronic equipment	3663
CPQ	Personal computers	3571
CY	Semiconductor—broad line	3674
TER	Semiconductor equip/materials	3825
NSM	Semiconductor—broad line	3674
HWP	Diversified computer systems	3570
IBM	Diversified computer systems	3570
	u^{997}	
PDG	Gold	1040
NEM	Gold	1040
NGC	Gold	1040
ABX	Gold	1040
ASA	Closed, end fund (gold)	6799
HM	Gold	1040
BMG	Gold	1040
AU	Gold	1040
HSM	General building materials	5210
MU	Semiconductor—memory chips	3674
	u^{996}	
NEM	Gold	1040
PDG	Gold	1040
ABX	Gold	1040
HM	Gold	1040
NGC	Gold	1040
ASA	Closed, end fund (gold)	6799
BMG	Gold	1040
CHL	Wireless communications	4813
CMB	Money center banks	6021
CCI	Money center banks	6021
	u^{995}	
TMX	Telecommunication services/foreign	4813
TV	Broadcasting—television	4833
MXF	Closed, end fund (Foreign)	6726
ICA	Heavy construction	1600
GTR	Heavy construction	1600
CTC	Telecom services/foreign	4813
PB	Beverages—soft drinks	2086
YPF	Independent oil and gas	2911
TXN	Semiconductor—broad line	3674
MU	Semiconductor—memory chips	3674
	u^{994}	
BAC	Money center banks	6021
CHL	Wireless communications	4813
BK	Money center banks	6022
CCI	Money center banks	6021
CMB	Money center banks	6021
BT	Money center banks	6022
JPM	Money center banks	6022
MEL	Regional—northeast banks	6021
NB	Money center banks	6021
WFC	Money center banks	6021
	u^{993}	
BP	Oil and gas equipment/services	2911
MOB	Oil and gas equipment/services	2911
SLB	Oil and gas equipment/services	1389
TX	Major integrated oil/gas	2911
UCL	Oil and gas refining/marketing	1311
ARC	Oil and gas equipment/services	2911
BHI	Oil and gas equipment/services	3533
CHV	Major integrated oil/gas	2911
APC	Independent oil and gas	1311
AN	Auto dealerships	2911
	u^{992}	
FPR	Auto manufacturers/major	3711
F	Auto manufacturers/major	3711
C	Auto manufacturers/major	3711
GM	Auto manufacturers/major	3711
TXN	Semiconductor—broad line	3674
ADI	Semiconductor—broad line	3674
CY	Semiconductor—broad line	3674
TER	Semiconductor equip/materials	3825
MGA	Auto parts	3714
LSI	Semiconductor—specialized	3674
	u^{991}	
ABT	Drug manufacturers/major	2834
PFE	Drug manufacturers/major	2834
SGP	Drug manufacturers/major	2834
LLY	Drug manufacturers/major	2834
JNJ	Drug manufacturers/major	2834
AHC	Oil and gas refining/marketing	2911
BMY	Drug manufacturers/major	2834
HAL	Oil and gas equipment/services	1600
WLA	Drug manufacturers/major	2834
BHI	Oil and gas equipment/services	3533

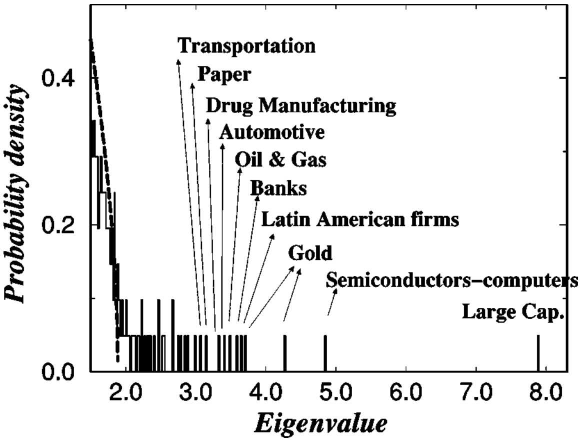

FIG. 13. Schematic illustration of the interpretation of the eigenvectors corresponding to the eigenvalues that deviate from the RMT upper bound. The dashed curve shows the RMT result of Eq. (6).

One eigenvector (u^{995}) displays a mixture of three industry groups—telecommunications, metal mining, and banking. An examination of these firms shows significant business activity in Latin America. Our results are also represented schematically in Fig. 13. A similar classification of stocks into sectors using different methods is obtained in Ref. [16].

Instead of performing the regression of Eq. (19), we remove the U-shaped intraday pattern using the procedure of Ref. [58] and compute C. The rationale behind this procedure is that, if two stocks are correlated, then the intraday pattern in volatility can give rise to weak intraday patterns in returns, which in turn affects the content of the eigenvectors. The results obtained by removing the intraday patterns are consistent with those obtained using the procedure of using the residuals of the regression of Eq. (19) to compute C (Table I). Often C is constructed from returns at longer time scales of $\Delta t = 1$ week or 1 month to avoid short-time scale effects [59].

E. Smallest eigenvalues and their corresponding eigenvectors

Having examined the largest eigenvalues, we next focus on the smallest eigenvalues which show large values of I^k [Fig. 11]. We find that the eigenvectors corresponding to the smallest eigenvalues contain as significant participants, pairs of stocks that have the largest values of C_{ij} in our sample. For example, the two largest components of u^1 correspond to the stocks of Texas Instruments (TXN) and Micron Technology (MU) with $C_{ij} = 0.64$, the largest correlation coefficient in our sample. The largest components of u^2 are Telefonos de Mexico (TMX) and Grupo Televisa (TV) with $C_{ij} = 0.59$ (second largest correlation coefficient). The eigenvector u^3 shows Newmont Gold Company (NGC) and Newmont Mining Corporation (NEM) with $C_{ij} = 0.50$ (third largest correlation coefficient) as largest components. In all three eigenvectors, the relative sign of the two largest components is *negative*. Thus pairs of stocks with a correlation coefficient much larger than the average $\langle C_{ij} \rangle$ effectively "decouple" from other stocks.

The appearance of strongly correlated pairs of stocks in the eigenvectors corresponding to the smallest eigenvalues of C can be qualitatively understood by considering the example of a 2×2 cross-correlation matrix

$$\mathsf{C}_{2\times 2} = \begin{bmatrix} 1 & c \\ c & 1 \end{bmatrix}. \tag{21}$$

The eigenvalues of $\mathsf{C}_{2\times 2}$ are $\beta_{\pm} = 1 \pm c$. The smaller eigenvalue β_{-} decreases monotonically with increasing cross-correlation coefficient c. The corresponding eigenvector is the antisymmetric linear combination of the basis vectors $\binom{1}{0}$ and $\binom{0}{1}$, in agreement with our empirical finding that the relative sign of largest components of eigenvectors corresponding to the smallest eigenvalues is negative. In this simple example, the symmetric linear combination of the two basis vectors appears as the eigenvector of the large eigenvalue β_{+}. Indeed, we find that TXN and MU are the largest components of u^{998}, TMX and TV are the largest components of u^{995}, and NEM and NGC are the largest and third largest components of u^{997}.

VII. STABILITY OF EIGENVECTORS IN TIME

We next investigate the degree of stability in time of the eigenvectors corresponding to the eigenvalues that deviate from RMT results. Since deviations from RMT results imply genuine correlations which remain stable in the period used to compute C, we expect the deviating eigenvectors to show some degree of time stability.

We first identify the p eigenvectors corresponding to the p largest eigenvalues which deviate from the RMT upper bound λ_{+}. We then construct a $p \times N$ matrix D with elements $D_{kj} = \{u_j^k; k=1,\ldots,p; j=1,\ldots,N\}$. Next, we compute a $p \times p$ "overlap matrix" $\mathsf{O}(t,\tau) = \mathsf{D}_A \mathsf{D}_B^T$, with elements O_{ij} defined as the scalar product of eigenvector u^i of period A (starting at time $t=t$) with u^j of period B at a later time $t+\tau$,

$$O_{ij}(t,\tau) \equiv \sum_{k=1}^{N} D_{ik}(t) D_{jk}(t+\tau). \tag{22}$$

If all the p eigenvectors are "perfectly" nonrandom and stable in time $O_{ij} = \delta_{ij}$.

We study the overlap matrices O using both high frequency and daily data. For high-frequency data ($L = 6448$ records at 30-min intervals), we use a moving window of length $L = 1612$, and slide it through the entire 2-yr period using discrete time steps $L/4 = 403$. We first identify the eigenvectors of the correlation matrices for each of these time periods. We then calculate overlap matrices $\mathsf{O}(t=0,\tau = nL/4)$, where $n \in \{1,2,3,\ldots\}$, between the eigenvectors for $t=0$ and for $t=\tau$.

Figure 14 shows a gray scale pixel representation of the matrix O (t,τ), for different τ. First, we note that the eigenvectors that deviate from RMT bounds show varying degrees of stability $[O_{ij}(t,\tau)]$ in time. In particular, the stability in time is largest for u^{1000}. Even at lags of $\tau = 1$ yr the corre-

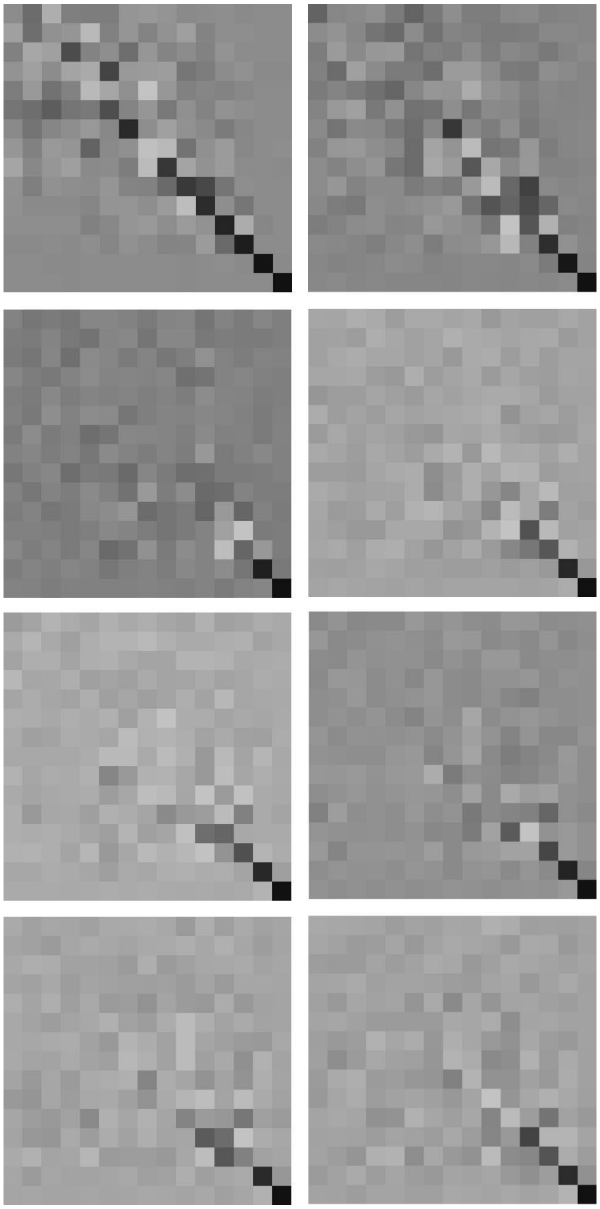

FIG. 14. Grayscale pixel representation of the overlap matrix $O(t,\tau)$ as a function of time for 30-min data for the 2-yr period 1994–1995. Here, the gray scale coding is such that black corresponds to $O_{ij}=1$ and white corresponds to $O_{ij}=0$. The length of the time window used to compute C is $L=1612$ (≈ 60 days) and the separation $\tau=L/4=403$ used to calculate successive O_{ij}. Thus, the left figure on the first row corresponds to the overlap between the eigenvector from the starting $t=0$ window and the eigenvector from time window $\tau=L/4$ later. The right figure is for $\tau=2L/4$. In the same way, the left figure on the second row is for $\tau=3L/4$, the right figure for $\tau=4L/4$, and so on. Even for large $\tau\approx 1$ yr, the largest four eigenvectors show large values of O_{ij}.

sponding overlap ≈ 0.85. The remaining eigenvectors show decreasing amounts of stability as the RMT upper bound λ_+ is approached. In particular, the three to four largest eigenvectors show large values of O_{ij} for up to $\tau=1$ yr.

Next, we repeat our analysis for daily returns of 422 stocks using 8685 records of 1-day returns, and a sliding window of length $L=965$ with discrete time steps $L/5 = 193$ days. Instead of calculating $O(t,\tau)$ for all starting points t, we calculate $O(\tau)\equiv\langle O(t,\tau)\rangle_t$, averaged over all $t = n\,L/5$, where $n\in\{0,1,2,\dots\}$. Figure 15 shows gray scale representations of $O(\tau)$ for increasing τ. We find similar results as found for shorter time scales, and find that eigenvectors corresponding to the largest two eigenvalues are stable for time scales as large as $\tau=20$ yr. In particular, the eigenvector u^{422} shows an overlap of ≈ 0.8 even over time scales of $\tau=30$ yr.

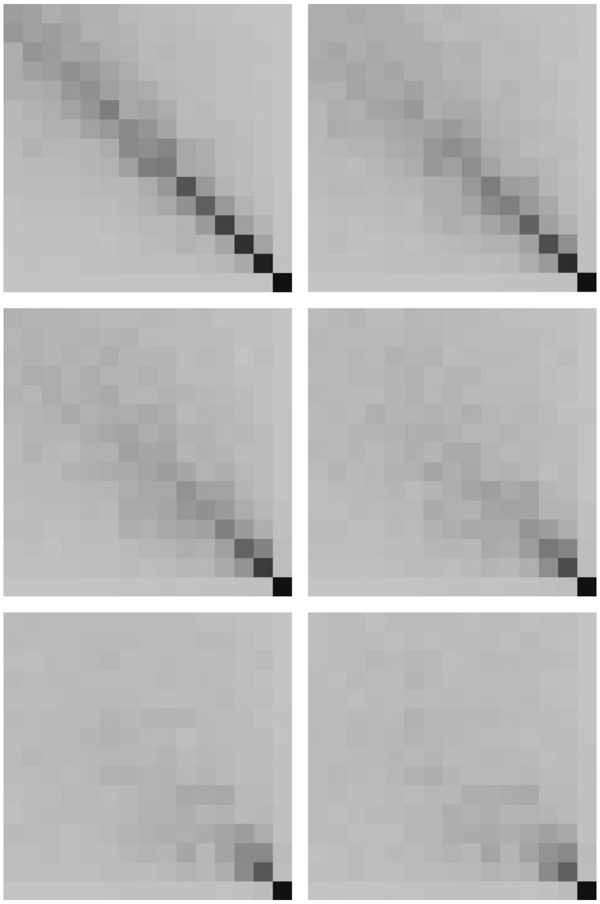

FIG. 15. Grayscale pixel representation of the overlap matrix $\langle O(t,\tau)\rangle_t$ for 1-day data, where we have averaged over all starting points t. Here, the length of the time window used to compute C is $L=965$ (≈ 4 yr) and the separation $\tau=L/5=193$ days used to calculate O_{ij}. Thus, the left figure on the first row is for $\tau=L/5$ and the right figure is for $\tau=2L/5$. In the same way, the left figure on the second row is for $\tau=3L/5$, the right figure for $\tau=4L/5$, and so on. Even for large $\tau\approx 20$ yr, the largest two eigenvectors show large values of O_{ij}.

VIII. APPLICATIONS TO PORTFOLIO OPTIMIZATION

The randomness of the "bulk" seen in the previous sections has implications in optimal portfolio selection [59]. We illustrate these using the Markowitz theory of optimal portfolio selection [3,17,60,61]. Consider a portfolio $\Pi(t)$ of stocks with prices S_i. The return on $\Pi(t)$ is given by

$$\Phi=\sum_{i=1}^{N} w_i G_i, \tag{23}$$

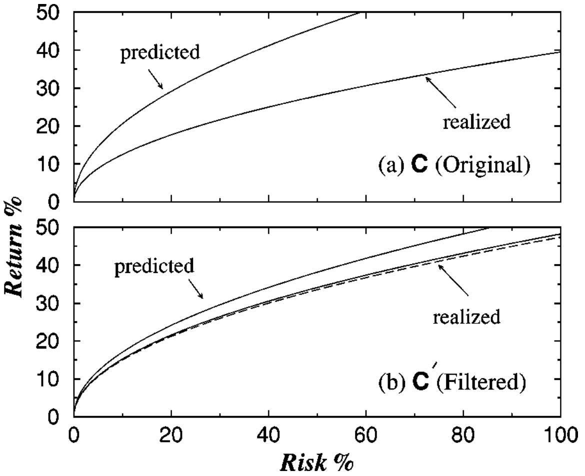

FIG. 16. (a) Portfolio return R as a function of risk D^2 for the family of optimal portfolios (without a risk-free asset) constructed from the original matrix C. The top curve shows the predicted risk D_{p}^2 in 1995 of the family of optimal portfolios for a given return, calculated using 30-min returns for 1995 and the correlation matrix C_{94} for 1994. For the same family of portfolios, the bottom curve shows the realized risk D_{r}^2 calculated using the correlation matrix C_{95} for 1995. These two curves differ by a factor of $D_{\mathrm{r}}^2/D_{\mathrm{p}}^2 \approx 2.7$. (b) Risk-return relationship for the optimal portfolios constructed using the filtered correlation matrix C′. The top curve shows the predicted risk D_{p}^2 in 1995 for the family of optimal portfolios for a given return, calculated using the filtered correlation matrix C'_{94}. The bottom curve shows the realized risk D_{r}^2 for the same family of portfolios computed using C'_{95}. The predicted risk is now closer to the realized risk: $D_{\mathrm{r}}^2/D_{\mathrm{p}}^2 \approx 1.25$. For the same family of optimal portfolios, the dashed curve shows the realized risk computed using the original correlation matrix C_{95} for which $D_{\mathrm{r}}^2/D_{\mathrm{p}}^2 \approx 1.3$.

where $G_i(t)$ is the return on stock i and w_i is the fraction of wealth invested in stock i. The fractions w_i are normalized such that $\Sigma_{i=1}^{N} w_i = 1$. The risk in holding the portfolio $\Pi(t)$ can be quantified by the variance

$$\Omega^2 = \sum_{i=1}^{N} \sum_{j=1}^{N} w_i w_j C_{ij} \sigma_i \sigma_j , \tag{24}$$

where σ_i is the standard deviation (average volatility) of G_i, and C_{ij} are elements of the cross-correlation matrix C. In order to find an optimal portfolio, we must minimize Ω^2 under the constraint that the return on the portfolio is some fixed value Φ. In addition, we also have the constraint that $\Sigma_{i=1}^{N} w_i = 1$. Minimizing Ω^2 subject to these two constraints can be implemented by using two Lagrange multipliers, which yields a system of linear equations for w_i, which can then be solved. The optimal portfolios thus chosen can be represented as a plot of the return Φ as a function of risk Ω^2 [Fig. 16].

To find the effect of randomness of C on the selected optimal portfolio, we first partition the time period 1994–1995 into two one-yr periods. Using the cross-correlation matrix C_{94} for 1994, and G_i for 1995, we construct a family of optimal portfolios, and plot Φ as a function of the predicted risk Ω_{p}^2 for 1995 [Fig. 16(a)]. For this family of portfolios, we also compute the risk Ω_{r}^2 realized during 1995 using C_{95} [Fig. 16(a)]. We find that the predicted risk is significantly smaller when compared to the realized risk,

$$\frac{\Omega_{\mathrm{r}}^2 - \Omega_{\mathrm{p}}^2}{\Omega_{\mathrm{p}}^2} \approx 170\%. \tag{25}$$

Since the meaningful information in C is contained in the deviating eigenvectors (whose eigenvalues are outside the RMT bounds), we must construct a "filtered" correlation matrix C′, by retaining only the deviating eigenvectors. To this end, we first construct a diagonal matrix Λ', with elements $\Lambda'_{ii} = \{0, \ldots, 0, \lambda_{988}, \ldots, \lambda_{1000}\}$. We then transform Λ' to the basis of C, thus obtaining the "filtered" cross-correlation matrix C′. In addition, we set the diagonal elements $C'_{ii} = 1$, to preserve $\mathrm{Tr}(\mathsf{C}) = \mathrm{Tr}(\mathsf{C}') = N$. We repeat the above calculations for finding the optimal portfolio using C′ instead of C in Eq. (24). Figure 16(b) shows that the realized risk is now much closer to the predicted risk

$$\frac{\Omega_{\mathrm{r}}^2 - \Omega_{\mathrm{p}}^2}{\Omega_{\mathrm{p}}^2} \approx 25\%. \tag{26}$$

Thus, the optimal portfolios constructed using C′ are significantly more stable in time.

IX. CONCLUSIONS

How can we understand the deviating eigenvalues, i.e., correlations that are stable in time? One approach is to postulate that returns can be separated into idiosyncratic and common components, i.e., that returns can be separated into different additive "factors," which represent various economic influences that are common to a set of stocks such as the type of industry, or the effect of news [4,36–54,62,63].

On the other hand, in physical systems one starts from the interactions between the constituents, and then relates interactions to correlated "modes" of the system. In economic systems, we ask if a similar mechanism can give rise to the correlated behavior. In order to answer this question, we model stock price dynamics by a family of stochastic differential equations [64], which describe the "instantaneous" returns $g_i(t) = (d/dt)\ln S_i(t)$ as a random walk with couplings J_{ij},

$$\tau_o \partial_t g_i(t) = -r_i g_i(t) - \kappa g_i^3(t) + \sum_j J_{ij} g_j(t) + \frac{1}{\tau_o} \xi_i(t). \tag{27}$$

Here, $\xi_i(t)$ are Gaussian random variables with correlation function $\langle \xi_i(t) \xi_j(t') \rangle = \delta_{ij} \tau_o \delta(t - t')$, and τ_o sets the time scale of the problem. In the context of a soft-spin model, the first two terms in the right-hand side of Eq. (27) arise from the derivative of a double-well potential, enforcing the soft-spin constraint. The interaction among soft spins is given by the couplings J_{ij}. In the absence of the cubic term, and without interactions, τ_o / r_i are relaxation times of the

$\langle g_i(t) g_i(t+\tau) \rangle$ correlation function. The return G_i at a finite time interval Δt is given by the integral of g_i over Δt.

Equation (27) is similar to the linearized description of interacting "soft spins" [65] and is a generalized case of the models of Ref. [64]. Without interactions, the variance of price changes on a scale $\Delta t \gg \tau_i$ is given by $\langle (G_i(\Delta t))^2 \rangle = \Delta t/(r^2 \tau_i)$, in agreement with recent studies [66], where stock price changes are described by an anomalous diffusion and the variance of price changes is decomposed into a product of trading frequency (analog of $1/\tau_i$) and the square of an "impact parameter" that is related to liquidity (analog of $1/r$).

As the coupling strengths increase, the soft-spin system undergoes a transition to an ordered state with permanent local magnetizations. At the transition point, the spin dynamics are very "slow" as reflected in a power-law decay of the spin autocorrelation function in time. To test whether this signature of strong interactions is present for the stock market problem, we analyze the correlation functions $c^{(k)}(\tau) \equiv \langle G^{(k)}(t) G^{(k)}(t+\tau) \rangle$, where $G^{(k)}(t) \equiv \Sigma_{i=1}^{1000} u_i^k G_i(t)$ is the time series defined by eigenvector u^k. Instead of analyzing $c^{(k)}(\tau)$ directly, we apply the detrended fluctuation analysis (DFA) method [67]. Figure 17 shows that the correlation functions $c^{(k)}(\tau)$ indeed decay as power laws [68] for the deviating eigenvectors u^k—in sharp contrast to the behavior of $c^{(k)}(\tau)$ for the rest of the eigenvectors and the autocorrelation functions of individual stocks, which show only short-ranged correlations. We interpret this as evidence for strong interactions [69].

In the absence of the nonlinearities (cubic term), we obtain only exponentially decaying correlation functions for the "modes" corresponding to the large eigenvalues, which is inconsistent with our finding of power-law correlations.

To summarize, we have tested the eigenvalue statistics of the empirically measured correlation matrix C against the null hypothesis of a random correlation matrix. This allows us to distinguish genuine correlations from "apparent" correlations that are present even for random matrices. We find that the bulk of the eigenvalue spectrum of C shares universal properties with the Gaussian orthogonal ensemble of random matrices. Further, we analyze the deviations from RMT, and find that (i) the largest eigenvalue and its corresponding eigenvector represent the influence of the entire market on all stocks, and (ii) using the rest of the deviating eigenvectors, we can partition the set of all stocks studied into distinct subsets whose identity corresponds to conventionally identified business sectors. These sectors are stable in time, in some cases for as many as 30 years. Finally, we have seen that the deviating eigenvectors are useful for the construction of optimal portfolios that have a stable ratio of risk to return.

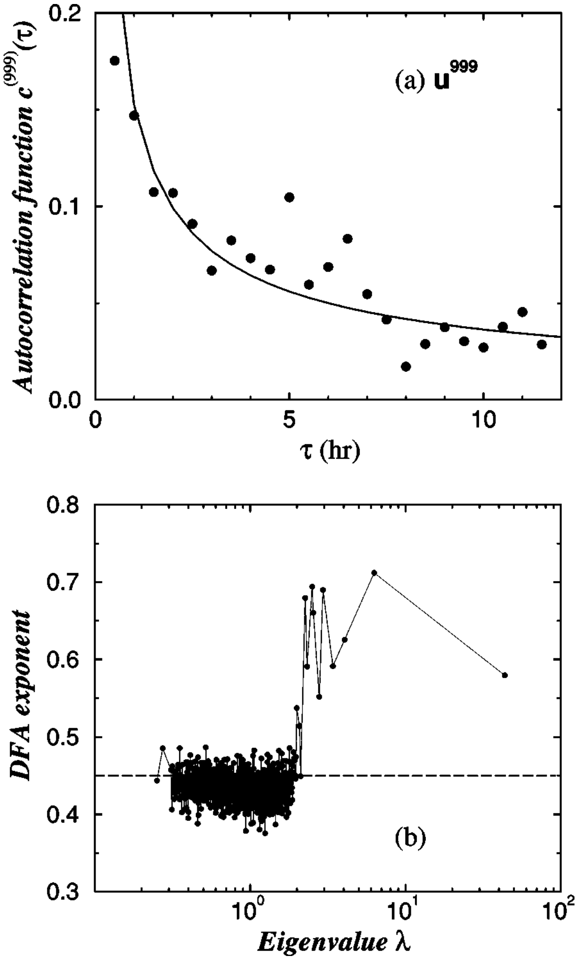

FIG. 17. (a) Autocorrelation function $c^{(k)}(\tau)$ of the time series defined by the eigenvector u^{999}. The solid line shows a fit to a power-law functional form $\tau^{-\gamma_k}$, whereby we obtain values $\gamma_k = 0.61 \pm 0.06$. (b) To quantify the exponents γ_k for all $k = 1, \ldots, 1000$ eigenvectors, we use the method of DFA analysis [66] often used to obtain accurate estimates of power-law correlations. We plot the detrended fluctuation function $F(\tau)$ as a function of the time scale τ for each of the 1000 time series. Absence of long-range correlations would imply $F(\tau) \sim \tau^{0.5}$, whereas $F(\tau) \sim \tau^{\nu}$ with $0.5 < \nu \leqslant 1$ implies power-law decay of the correlation function with exponent $\gamma = 2 - 2\nu$. We plot the exponents ν as a function of the eigenvalue and find values exponents ν significantly larger than 0.5 for all the deviating eigenvectors. In contrast, for the remainder of the eigenvectors, we obtain the mean value $\nu = 0.44 \pm 0.04$, comparable to the value $\nu = 0.5$ for the uncorrelated case.

ACKNOWLEDGMENTS

We thank J.-P. Bouchaud, P. Cizeau, E. Derman, X. Gabaix, J. Hill, L. Viciera, and J. Z. Zou for helpful discussions. We thank O. Bohigas for pointing out Ref. [26] to us. B.R. thanks DFG Grant No. RO1-1/2447 for financial support. T.G. thanks Boston University for warm hospitality. The Center for Polymer Studies is supported by the NSF, British Petroleum, the NIH, and the National Center for Research Resources (Grant No. P41 RR13622).

APPENDIX: "UNFOLDING" THE EIGENVALUE DISTRIBUTION

As discussed in Sec. V, random matrices display *universal* functional forms for eigenvalue correlations that depend only on the general symmetries of the matrix. A first step to test the data for such universal properties is to find a transformation called "unfolding," which maps the eigenvalues λ_i to new variables called ''unfolded eigenvalues'' ξ_i, whose distribution is uniform [11–13]. Unfolding ensures that the distances between eigenvalues are expressed in units of *local* mean eigenvalue spacing [11], and thus facilitates comparison with analytical results.

We first define the cumulative distribution function of eigenvalues, which counts the number of eigenvalues in the interval $\lambda_i \leqslant \lambda$,

$$F(\lambda) = N \int_{-\infty}^{\lambda} P(x)\,dx, \tag{A1}$$

where $P(x)$ denotes the probability density of eigenvalues and N is the total number of eigenvalues. The function $F(\lambda)$ can be decomposed into an average and a fluctuating part,

$$F(\lambda) = F_{\mathrm{av}}(\lambda) + F_{\mathrm{fluc}}(\lambda). \tag{A2}$$

Since $P_{\mathrm{fluc}} \equiv dF_{\mathrm{fluc}}(\lambda)/d\lambda = 0$ on average,

$$P_{\mathrm{rm}}(\lambda) \equiv \frac{dF_{\mathrm{av}}(\lambda)}{d\lambda} \tag{A3}$$

is the averaged eigenvalue density. The dimensionless, unfolded eigenvalues are then given by

$$\xi_i \equiv F_{\mathrm{av}}(\lambda_i). \tag{A4}$$

Thus, the problem is to find $F_{\mathrm{av}}(\lambda)$. We follow two procedures for obtaining the unfolded eigenvalues ξ_i: (i) a phenomenological procedure referred to as Gaussian broadening [11–13], and (ii) fitting the cumulative distribution function $F(\lambda)$ of Eq. (A1) with the analytical expression for $F(\lambda)$ using Eq. (6). These procedures are discussed below.

1. Gaussian broadening

Gaussian broadening [70] is a phenomenological procedure that aims at approximating the function $F_{\mathrm{av}}(\lambda)$ defined in Eq. (A2) using a series of Gaussian functions. Consider the eigenvalue distribution $P(\lambda)$, which can be expressed as

$$P(\lambda) = \frac{1}{N} \sum_{i=1}^{N} \delta(\lambda - \lambda_i). \tag{A5}$$

The δ functions about each eigenvalue are approximated by choosing a Gaussian distribution centered around each eigenvalue with standard deviation $(\lambda_{k+a} - \lambda_{k-a})/2$, where $2a$ is the size of the window used for broadening [71]. Integrating Eq. (A5) provides an approximation to the function $F_{\mathrm{av}}(\lambda)$ in the form of a series of error functions, which using Eq. (A4) yields the unfolded eigenvalues.

2. Fitting the eigenvalue distribution

Phenomenological procedures are likely to contain artificial scales, which can lead to an ''overfitting'' of the smooth part $F_{\mathrm{av}}(\lambda)$ by adding contributions from the fluctuating part $F_{\mathrm{fluc}}(\lambda)$. The second procedure for unfolding aims at circumventing this problem by fitting the cumulative distribution of eigenvalues $F(\lambda)$ [Eq. (A1)] with the analytical expression for

$$F_{\mathrm{rm}}(\lambda) = N \int_{-\infty}^{\lambda} P_{\mathrm{rm}}(x)\,dx, \tag{A6}$$

where $P_{\mathrm{rm}}(\lambda)$ is the probability density of eigenvalues from Eq. (6). The fit is performed with λ_-, λ_+, and N as free parameters. The fitted function is an estimate for $F_{\mathrm{av}}(\lambda)$, whereby we obtain the unfolded eigenvalues ξ_i. One difficulty with this method is that the deviations of the spectrum of C from Eq. (6) can be quite pronounced in certain periods, and it is difficult to find a good fit of the cumulative distribution of eigenvalues to Eq. (A6).

[1] J.D. Farmer, Comput. Sci. Eng. **1**, 26 (1999).

[2] R. N. Mantegna and H. E. Stanley, *An Introduction to Econophysics: Correlations and Complexity in Finance* (Cambridge University Press, Cambridge, 2000).

[3] J. P. Bouchaud and M. Potters, *Theory of Financial Risk* (Cambridge University Press, Cambridge 2000).

[4] J. Campbell, A. W. Lo, and A. C. MacKinlay, *The Econometrics of Financial Markets* (Princeton University Press, Princeton, 1997).

[5] E.P. Wigner, Ann. Math. **53**, 36 (1951); Proc. Cambridge Philos. Soc. **47**, 790 (1951).

[6] E. P. Wigner, in *Conference on Neutron Physics by Time-of-flight* (Oak Ridge National Laboratories Press, Gatlinburg, Tennessee, 1956), pp. 59.

[7] E.P. Wigner, Proc. Cambridge Philos. Soc. **47**, 790 (1951).

[8] F.J. Dyson, J. Math. Phys. **3**, 140 (1962).

[9] F.J. Dyson and M.L. Mehta, J. Math. Phys. **4**, 701 (1963).

[10] M.L. Mehta and F.J. Dyson, J. Math. Phys. **4**, 713 (1963).

[11] M. L. Mehta, *Random Matrices* (Academic Press, Boston, 1991).

[12] T.A. Brody, J. Flores, J.B. French, P.A. Mello, A. Pandey, and S.S.M. Wong, Rev. Mod. Phys. **53**, 385 (1981).

[13] T. Guhr, A. Müller-Groeling, and H.A. Weidenmüller, Phys. Rep. **299**, 190 (1998).

[14] L. Laloux, P. Cizeau, J.-P. Bouchaud, and M. Potters, Phys. Rev. Lett. **83**, 1467 (1999).

[15] V. Plerou, P. Gopikrishnan, B. Rosenow, L.A.N. Amaral, and H.E. Stanley, Phys. Rev. Lett. **83**, 1471 (1999).

[16] R.N. Mantegna, Eur. Phys. J. B **11**, 193 (1999); L. Kullmann, J. Kertész, and R.N. Mantegna, e-print cond-mat/0002238; an interesting analysis of cross correlation between stock market indices can be found in G. Bonanno, N. Vandewalle, and R.N. Mantegna, Phys. Rev. E **62**, 7615 (2000).

[17] P. Gopikrishnan, B. Rosenow, V. Plerou, and H.E. Stanley, Phys. Rev. E **64**, 035106(R) (2001).

[18] P. Ormerod and C. Mounfield, Physica A **280**, 497 (2000). Use

RMT methods to analyze cross correlations in the gross domestic product (GDP) of countries.

[19] The time series of prices have been adjusted for stock splits and dividends.

[20] Only those stocks, which have survived the 2-yr period 1994–1995 were considered in our analysis.

[21] V. Plerou, P. Gopikrishnan, L.A.N. Amaral, M. Meyer, and H.E. Stanley, Phys. Rev. E **60**, 6519 (1999); P. Gopikrishnan, V. Plerou, L.A.N. Amaral, M. Meyer, and H.E. Stanley, *ibid.* **60**, 5305 (1999); P. Gopikrishnan, M. Meyer, L.A.N. Amaral, and H.E. Stanley, Eur. Phys. J. B **3**, 139 (1998).

[22] T. Lux, Appl. Financ. Econ. **6**, 463 (1996).

[23] M. Lauretan, Global Investor **135**, 65 (2000).

[24] P. Silvapulle and C.W.J. Granger, Quantitative Finance **1**, 542 (2001).

[25] A.K. Tsui and Q. Yu, Math. Comput. Simul. **48**, 503 (1999).

[26] F.J. Dyson, Rev. Mex. Fis. **20**, 231 (1971).

[27] A.M. Sengupta and P.P. Mitra, Phys. Rev. E **60**, 3389 (1999).

[28] M.J. Bowick and E. Brézin, Phys. Lett. B **268**, 21 (1991); J. Feinberg and A. Zee, J. Stat. Phys. **87**, 473 (1997).

[29] For the case of Lévy random matrices applied to financial covariances, see Z. Burda *et al.*, e-print cond-mat/0103140; e-print cond-mat/0103109; e-print cond-mat/0103108.

[30] To perform diagonalization of large correlation matrices, we first reduce the correlation matrix to tridiagonal form, which is computationally five times more efficient than direct diagonalization. To obtain accurate estimates, we use the QL algorithm with implicit shifts that is known to work ‘‘extremely well in practice’’ [W. H. Press, *et al.*, *Numerical Recipes*, 2nd ed. (Cambridge University Press, Cambridge, 1999)].

[31] Analytical evidence for this ‘‘universality’’ is summarized in Sec. VIII of Ref. [13].

[32] A recent review is J.J.M. Verbaarschot and T. Wettig, Annu. Rev. Nucl. Part. Sci. **50**, 343 (2000); e-print hep-ph/0003017.

[33] For GOE matrices, the Wigner surmise [Eq. (9)] is not exact [11]. Despite this fact, the Wigner surmise is widely used because the difference between the exact form of $P_{\rm GOE}(s)$ and Eq. (9) is almost negligible [11].

[34] Analogous to the GOE, which is defined on the space of real symmetric matrices, one can define two other ensembles [11]: (i) the Gaussian unitary ensemble (GUE), which is defined on the space of Hermitian matrices, with the requirement that the joint probability of elements is invariant under unitary transformations, and (ii) the Gaussian symplectic ensemble (GSE), which is defined on the space of Hermitian ‘‘self-dual’’ matrices with the requirement that the joint probability of elements is invariant under symplectic transformations. Formal definitions can be found in Ref. [11].

[35] S. Drozdz, F. Gruemmer, F. Ruf, and J. Speth, e-print cond-mat/9911168.

[36] W. Sharpe, *Portfolio Theory and Capital Markets* (McGraw-Hill, New York, 1970).

[37] W. Sharpe, G. Alexander, and J. Bailey, *Investments*, 5th ed. (Prentice-Hall, Englewood Cliffs, 1995).

[38] W. Sharpe, J. Financ. **19**, 425 (1964).

[39] J. Lintner, Photogramm. Rev. Econ. Stat. **47**, 13 (1965).

[40] S. Ross, J. Econ. Theory **13**, 341 (1976).

[41] S. Brown and M. Weinstein, J. Financ. Econ. **14**, 491 (1985).

[42] F. Black, J. Business **45**, 444 (1972).

[43] M. Blume and I. Friend, J. Financ. **28**, 19 (1973).

[44] E. Fama and K. French, J. Financ. **47**, 427 (1992); J. Financ. Econ. **33**, 3 (1993).

[45] E. Fama and J. Macbeth, J. Political Econ. **71**, 607 (1973).

[46] R. Roll and S. Ross, J. Financ. **49**, 101 (1994).

[47] N. Chen, R. Roll, and S. Ross, J. Business **59**, 383 (1986).

[48] R.C. Merton, Econometrica **41**, 867 (1973).

[49] B. Lehmann and D. Modest, J. Financ. Econ. **21**, 213 (1988).

[50] J. Campbell, J. Political Econ. **104**, 298 (1996).

[51] J. Campbell and J. Ammer, J. Financ. **48**, 3 (1993).

[52] G. Connor and R. Korajczyk, J. Financ. Econ. **15**, 373 (1986); **21**, 255 (1988); J. Financ. **48**, 1263 (1993).

[53] A non-Gaussian one-factor model and its relevance to cross correlations is investigated in P. Cizeau, M. Potters, and J.-P. Bouchaud, e-print cond-mat/0006034.

[54] The limitations of a one-factor description as regards extreme market fluctuations can be found in F. Lillo and R.N. Mantegna, Phys. Rev. E **62**, 6126 (2000); e-print cond-mat/0002438.

[55] Y.V. Fyodorov and A.D. Mirlin, Phys. Rev. Lett. **69**, 1093 (1992); **71**, 412 (1993); Int. J. Mod. Phys. B **8**, 3795 (1994); A.D. Mirlin and Y.V. Fyodorov, J. Phys. A **26**, L551 (1993); E.P. Wigner, Ann. Math. **62**, 548 (1955).

[56] P.A. Lee and T.V. Ramakrishnan, Rev. Mod. Phys. **57**, 287 (1985).

[57] Metals or semiconductors with impurities can be described by Hamiltonians with random-hopping integrals [F. Wegner and R. Oppermann, Z. Phys. B **34**, 327 (1979)]. Electron-hopping between neighboring sites is more probable than hopping over large distances, leading to a Hamiltonian that is a random band matrix.

[58] Y. Liu, P. Gopikrishnan, P. Cizeau, M. Meyer, C.-K. Peng, and H.E. Stanley, Phys. Rev. E **60**, 1390 (1999); Y. Liu, P. Cizeau, M. Meyer, C.-K. Peng, and H.E. Stanley, Physica A **245**, 437 (1997); P. Cizeau, Y. Liu, M. Meyer, C.-K. Peng, and H.E. Stanley, *ibid.* **245**, 441 (1997).

[59] E. J. Elton and M. J. Gruber, *Modern Portfolio Theory and Investment Analysis* (Wiley, New York, 1995).

[60] L. Laloux *et al.*, Int. J. Theor. Appl. Finance **3**, 391 (2000).

[61] F. Black and R. Litterman, Financial Analysts J. **28**, (1992).

[62] J.D. Noh, Phys. Rev. E **61**, 5981 (2000).

[63] M. Marsili, e-print cond-mat/0003241.

[64] J.D. Farmer, e-print adap-org/9812005; R. Cont and J.-P. Bouchaud, Eur. Phys. J. B **6**, 543 (1998).

[65] K. H. Fischer and J. A. Hertz, *Spin Glasses* (Cambridge University Press, New York, 1991).

[66] V. Plerou, P. Gopikrishnan, L.A.N. Amaral, X. Gabaix, and H.E. Stanley, Phys. Rev. E **62**, R3023 (2000).

[67] C.K. Peng *et al.*, Phys. Rev. E **49**, 1685 (1994).

[68] In contrast, the autocorrelation function for the S&P 500 index returns $G_{\rm SP}(t)$ displays only correlations on short-time scales of <30 min, beyond which the autocorrelation function is at the level of noise [58]. On the other hand, the returns for individual stocks have pronounced anticorrelations on short-time scales (≈ 30 min), which is an effect of the bid-ask bounce [4]. For certain portfolios of stocks, returns are found to have long memory [A. Lo, Econometrica **59**, 1279 (1991)].

[69] For the case of predominantly ‘‘ferromagnetic’’ couplings ($J_{ij}>0$) within disjoint groups of stocks, a factor model [such as Eq. (19)] can be derived from the model of ‘‘interacting

stocks'' in Eq. (27). In the spirit of a mean-field approximation, the influence of the price changes of all other stocks in a group on the price of a given stock can be modeled by an effective field, which has to be calculated self-consistently. This effective field would then play a similar role as a factor in standard economic models.

[70] M. Brack, J. Damgaard, A.S. Jensen, H.C. Pauli, V.M. Strutinsky, and C.Y. Wong, Rev. Mod. Phys. **44**, 320 (1972).

[71] H. Bruus and J.-C. Anglés d'Auriac, Europhys. Lett. **35**, 321 (1996).