Identifying Business Sectors from Stock Price Fluctuations

Parameswaran Gopikrishnan[1], Bernd Rosenow[1,2], Vasiliki Plerou[1,3], and H. Eugene Stanley[1]
[1] *Center for Polymer Studies and Department of Physics, Boston University, Boston, Massachusetts 02215, USA*
[2] *Department of Physics, Harvard University, Cambridge, Massachusetts 02138, USA*
[3] *Department of Physics, Boston College, Chestnut Hill, Massachusetts 02167, USA*
(May 20, 2001)

Firms having similar business activities are correlated. We analyze two different cross–correlation matrices C constructed from (i) 30–min price fluctuations of 1000 US stocks for the 2–year period 1994–95 and (ii) 1–day price fluctuations of 422 US stocks for the 35–year period 1962–96. We find that the eigenvectors of C corresponding to the largest eigenvalues allow us to partition the set of all stocks into distinct subsets. These subsets are similar to conventionally-identified business sectors, and are stable for extended periods of time. Using a set of coupled stochastic differential equations, we argue how correlations between stocks might arise. Finally, we demonstrate that the sectors we identify are useful for the practical goal of finding an investment which earns a given return without exposure to unnecessary risk.

PACS numbers: 05.45.Tp, 89.90.+n, 05.40.-a, 05.40.Fb

The internal structure of a complex system manifests itself in correlations among its constituents. In physical systems, one relates correlations to basic interactions, but for the stock market problem [1], the underlying 'interactions' are not known. Suppose that the change of stock prices can be visualized by the motion of point particles. Correlated particle motion can be pictured as "strings" connecting pairs of particles. Given only the records of the particle positions at equal time intervals, how can we identify the strings without 'seeing' them? One approach is to first calculate the cross-correlation matrix C whose elements C_{ij} are the correlation-coefficients between the velocities of two particles i and j. The eigenvectors of C convey information about the collective modes of the system.

What is the analog of the cross-correlation matrix C for the stock market problem? We define the cross–correlation matrix C with elements $C_{ij} \equiv [\langle G_i G_j\rangle - \langle G_i\rangle\langle G_j\rangle]/\sigma_i\sigma_j$, where σ_i is the standard deviation of price fluctuations $G_i(t) \equiv \ln S_i(t+\Delta t) - \ln S_i(t)$ (returns), $S_i(t)$ denotes the price of stock $i = 1, \ldots, N$, and $\langle \ldots \rangle$ denotes a time average over the period studied. To investigate correlations on different time scales, we analyze (i) 30-min returns of $N = 1000$ largest stocks for the two-year period 1994-95 and (ii) daily returns of $N = 422$ stocks for the 35-year period 1962-96 [2].

We first diagonalize C and rank-order its eigenvalues λ_k such that $\lambda_{k+1} > \lambda_k$; the corresponding eigenvectors are denoted u^k. Next, we analyze the components of those *deviating eigenvectors* whose eigenvalues are larger than the upper bound for uncorrelated time series [3,4]. 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 Tr $\mathsf{C} = N$. Thus, in order to analyze the contents of the deviating eigenvectors, we first remove the effect of the largest eigenvalue [5].

To analyze the information contained in the eigenvectors u^k, we partition the 1000 stocks into groups labeled $\ell = 1 \ldots, 75$ (comprising N_ℓ stocks each) according to the first two digits of their Standard Industrial Classification (SIC) code, which classifies major industry groups. We define a projection matrix P, with elements $P_{\ell i} = 1/N_\ell$ if stock i belongs to group ℓ and $P_{\ell i} = 0$ otherwise. For each deviating eigenvector u^k, we compute the contribution $X_\ell^k \equiv \sum_{i=1}^{N} P_{\ell i}^k [u_i^k]^2$ of each industry group ℓ. The above procedure of computing X_ℓ^k is analogous to the analysis of wave functions in disordered systems, where one calculates the probability of finding a particle in a given region.

Figure 1 shows X_ℓ^k for ten largest eigenvectors after excluding the influence of the largest eigenvalue. The contribution X_ℓ^{999} shows several industries. We examine the significant contributors and find mainly stocks with large market capitalization [Fig. 2]. We analyze X_ℓ^k for the remainder of the deviating eigenvectors and find a significant 'peak' at distinct values of the SIC code — suggesting that these eigenvectors correspond to distinct industry groups [6].

One deviating eigenvector u^{995} displays large values of X_ℓ^k for firms belonging to the heavy construction industry and telecommunications industry. In addition, an examination of these firms shows significant business activity in Latin America. Another case corresponds to eigenvectors u^{996} and u^{997}, both of which contain a mixture of stocks of gold-mining firms and banking firms. We find that these two sectors separate when we compute the symmetric and antisymmetric combinations $1/\sqrt{2}(\mathsf{u}^{996} \pm \mathsf{u}^{997})$. The remainder of the deviating eigenvectors display technology, metal mining, banking, petroleum refining, auto manufacturing, drug manufacturing, and paper manufacturing firms [Fig. 1].

We next focus on the interpretation of the largest eigenvalue λ_{1000}. Using the eigenvector u^{1000}, we construct a time series $G^{1000}(t) \equiv \sum_{i=1}^{1000} u_i^{1000} G_i(t)$. We then compare $G^{1000}(t)$ with the returns $G_{SP}(t)$ of the S&P 500 index, a benchmark for gauging the performance of entire US stock market. Regressing $G^{1000}(t)$ against $G_{SP}(t)$ shows a scatter around a linear fit with slope 0.85 ± 0.09 [Fig. 3]. Thus, we interpret the eigenvector u^{1000} as the influence of the entire market, that is common for all stocks [3,4].

Next, we examine whether the eigenvectors u^k corresponding to business sectors remain stable in time. Partitioning the year 1994 into two 6-month periods, A and B, we calculate the corresponding eigenvectors u_A and u_B of the cross-correlation matrices and quantify the time stability by calculating the magnitude of the scalar products $O_{ij} \equiv |\, \mathsf{u}_A^i \; \mathsf{u}_B^j \,|$ for the 20 largest eigenvalues. Perfect time stability would mean $O_{ij} = \delta_{ij}$. For $i = 1000$, we find $O_{ii} = 0.93$ — indicating almost perfect stability. We find that O_{ii} decreases as i decreases from 1000 [Fig. 4]. Extending this analysis to daily returns using database (ii) shows that the eigenvectors corresponding to the largest 3 eigenvalues are stable for as many as 10 years.

How can we explain correlations that are stable in time? 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 "interactions" give rise to the correlated behavior. Interactions can arise when two companies are doing business together, or compete for the same market. To study if the correlations can be explained through interactions [7], we model stock price dynamics by a differential equation [8,9], which describes the 'instantaneous" returns $g_i(t) = \frac{d}{dt} \ln S_i(t)$ as a random walk with interactions J_{ij}

$$\tau_i \partial_t g_i(t) = -r g_i(t) + \sum_j J_{ij} g_j(t) + \frac{1}{\tau_i} \xi_i(t) \quad . \qquad (1)$$

Here, $\xi_i(t)$ are Gaussian random variables with correlation function $\langle \xi_i(t) \xi_j(t') \rangle = \delta_{ij} \tau_i \delta(t - t')$, and τ_i are relaxation times of the $\langle g_i(t) g_i(t+T) \rangle$ correlation function. The return G_i at a finite time interval Δt is given by the integral of g_i over Δt.

Calculating time-dependent correlation functions for the g_i, we find that correlations caused by interactions are accompanied by a phenomenon analogous to "critical slowing down". The market time series $\mathrm{G}^{1000}(t)$ — as well as time series constructed similarly for other deviating eigenvectors — have considerably larger correlation times than a time series constructed out of a random eigenvector, consistent with the hypothesis that correlations between firms are caused by interactions.

The eigenvectors that we interpret as defining business sectors also have relevance to the practical goal of finding an investment which earns a given return without exposure to unnecessary risk ("optimal portfolio"). Risk can be reduced by diversification of investment into independently fluctuating groups of stocks, such as the mutually uncorrelated business sectors that we find. Since the sectors (eigenvectors) are stable in time, we expect the ratio of risk to return of the portfolios constructed from them to be stable.

Consider a portfolio $P(t) \equiv \sum_{i=1}^{N} w_i S_i(t)$, where w_i is the fraction of wealth invested in stock i. The portfolio return is given by $R = \sum_{i=1}^{N} w_i G_i$. The risk in holding the portfolio $P(t)$ can be quantified by the variance $D^2 = \sum_{i=1}^{N} \sum_{j=1}^{N} w_i w_j \, C_{ij} \, \sigma_i \sigma_j$, where σ_i is the standard deviation of G_i [10]. In order to find an optimal portfolio, we minimize D^2 under the constraints that the portfolio return is some fixed value R and $\sum_{i=1}^{N} w_i = 1$. We thereby obtain a family of optimal portfolios, which we represent by plotting R as a function of risk D^2 [Fig. 5].

To find the effect of randomness of the C_{ij} on optimal portfolio selection, we partition the time period 1994-95 into two 1-year periods. Using the cross-correlation matrix C_{94} for 1994, and G_i for 1995 [11], we construct a family of optimal portfolios and plot R as a function of the predicted risk D_{p}^2 for 1995 [Fig. 5(a)]. For this family of portfolios, we also compute the risk D_{r}^2 realized during 1995 using C_{95} [Fig. 5(a)]. We find that the predicted risk is significantly smaller than the realized risk: $[D_{\mathrm{r}}^2 - D_{\mathrm{p}}^2]/D_{\mathrm{p}}^2 \approx 170\%$.

Since the meaningful information in C is contained in the deviating eigenvectors that define business sectors, we construct a 'filtered' correlation matrix C', by retaining only the deviating eigenvectors [12]. We repeat the above calculations for finding the optimal portfolio using C' instead of C. Figure 5(b) shows that the realized risk is now much closer to the predicted risk: $[D_{\mathrm{r}}^2 - D_{\mathrm{p}}^2]/D_{\mathrm{p}}^2 \approx 25\%$. Thus, the optimal portfolios constructed using C' are significantly more stable in time.

In summary, given only the change in price of a stock, and no additional information about the stock, we can partition the set of all 10^3 stocks studied into subsets whose identities correspond well to conventionally-identified sectors of economic activity. The sector correlations are stable in time and can be used for the construction of optimal portfolios with a stable ratio of risk to return.

We thank J.-P. Bouchaud, P. Cizeau, E. Derman, X. Gabaix, J. Hill, M. Janjusevic, R. N. Mantegna, M. Potters, L. Viceira, J. Zou, and especially L. A. N. Amaral for stimulating discussions, and DFG grant RO1-1/2447 for financial support. Our results on the applications to portfolio selection were presented at the APS March 2000 meeting by one of us (BR) and independently by P. Cizeau.

[1] A recent review of research done by physicists is J. D. Farmer, Computing in Sci. & Eng. **1**, 26 (1999).

[2] The 30-min data analyzed are from the Trades and Quotes database published by the New York Stock Exchange. The daily data is from the Center for Research in Securities Prices (CRSP) database. In both cases, only those companies that survive the entire period are considered in our analysis.
[3] L. Laloux, *et al.*, Phys. Rev. Lett. **83**, 1469 (1999).
[4] V. Plerou, *et al.*, Phys. Rev. Lett. **83**, 1471 (1999).
[5] The presence of a very large eigenvalue, by preservation of the trace Tr (C)$= N$, tends to cause the other large eigenvalues and eigenvectors to be influenced by randomness. Therefore, in order to remove the effect of the "market mode" on other large eigenvalues, we implement the linear regression $G_i(t) = \alpha_i + \beta_i G^{1000}(t) + \epsilon_i(t)$, where $G^{1000} \equiv \sum_{i=1}^{1000} u_i^{1000}\, G_i(t)$. We then recompute the correlation matrix C using the residuals $\epsilon_i(t)$. Alternatively, removing the intra-daily patterns in absolute values of $G_i(t)$ using the procedure of Y. Liu *et al.*, Phys. Rev. E **60**, 1390 (1999) also yields similar results.
[6] An interesting classification of stocks into sectors using the concept of ultrametric distances was obtained by R. N. Mantegna, Eur. Phys. J. B **11**, 193 (1999).
[7] Equation (1) is similar to the linearized description of interacting soft spins. Unlike the spin model, different stocks have relaxation times which, in general, increase by two orders of magnitude with market capitalization — from 10 s up to 1000 s.
[8] J. D. Farmer, e-print adap-org/9812005; R. Cont and J.-P. Bouchaud, Eur. Phys. J. B **6**, 543 (1998).
[9] Equation (1) is a generalized case of the models of Refs. [8]. Without interactions, the variance of price changes on a scale $\Delta t \gg \tau_i$ is given by $\langle (G_i(\Delta t))^2 \rangle = \Delta t/(r^2 \tau_i)$. This is in agreement with recent studies [V. Plerou *et al.*, Phys. Rev. E. **62**, R3023 (Sept 00)], 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 (analog of $1/r$), which is related to liquidity.
[10] E. J. Elton and M. J. Gruber, *Modern Portfolio Theory and Investment Analysis* (J. Wiley, New York, 1995).
[11] We thank J.-P. Bouchaud for this suggestion.
[12] To filter C from the effects of the random part, 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 $\mathrm{Tr}(\mathsf{C}) = \mathrm{Tr}(\mathsf{C}') = N$.
[13] The curves specify $\Delta t = 30$ min risk and return values.

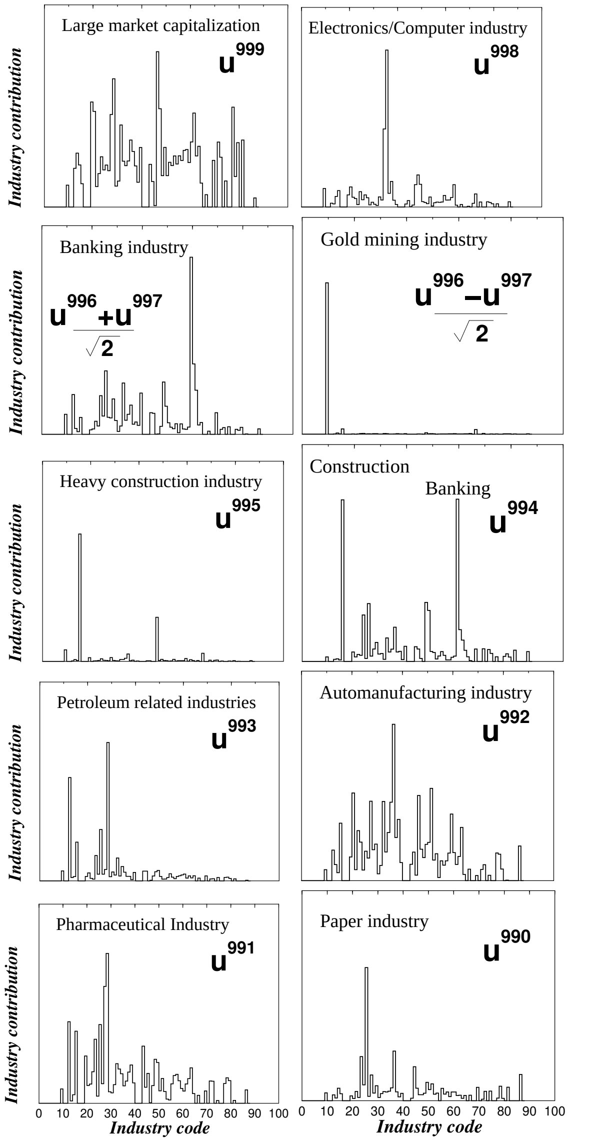

FIG. 1. Contribution X_ℓ^k to industry sector ℓ of eigenvector u^k for the deviating eigenvectors shows marked peaks at distinct values of SIC code, for all but u^{999} which contains stocks with large capitalizations as significant contributors.

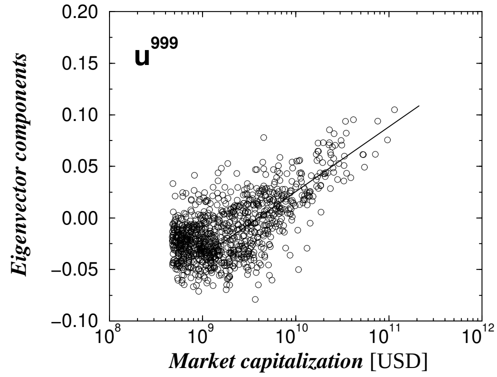

FIG. 2. All 10^3 eigenvector components of u^{999} plotted against market capitalization (in units of US Dollars) shows that large firms contribute more than small firms. The straight line, which shows a logarithmic fit, is a guide to the eye.

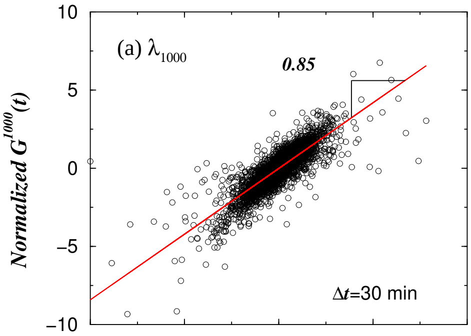

FIG. 3. S&P 500 returns $G_{SP}(t)$ regressed against the return $G^{1000}(t)$ of the portfolio defined by the eigenvector u^{1000}. Both axes are scaled by their respective standard deviations. A linear regression yields a slope 0.85 ± 0.09, showing a large degree of correlation.

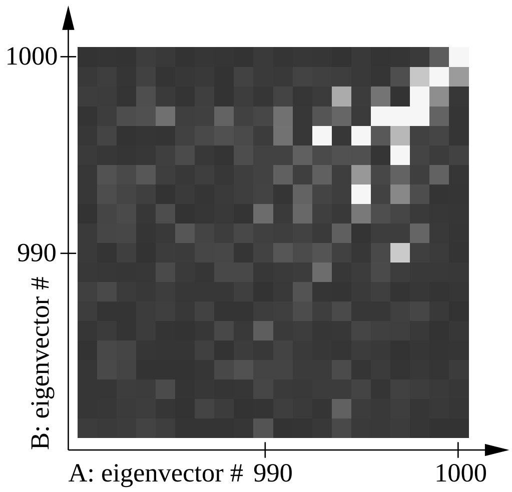

FIG. 4. Comparison of eigenvectors for different time periods A (first half of 1994) and B (second half of 1994) by means of their scalar product O_{ij}, represented on a greyscale, where zero (black) corresponds to no overlap, and white (one) to perfect overlap. Note that the eigenvectors corresponding to the 4 largest eigenvalues have a large degree of time stability.

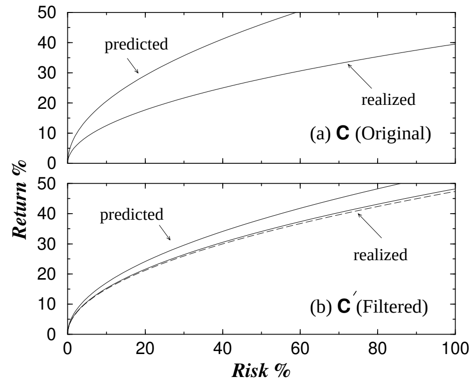

FIG. 5. (a) Portfolio return R as a function of risk D^2 for the family of optimal portfolios (without a risk-free asset) constructed from the original matrix C [13]. The top curve shows the predicted risk $D_{\rm p}^2$ in 1995 of the family of optimal portfolios for a given return, calculated using 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 $D_{\rm r}^2$ calculated using the correlation matrix C_{95} for 1995. These two curves differ by a factor of $D_{\rm r}^2/D_{\rm p}^2 \approx 2.7$. (b) Risk-return relationship for the optimal portfolios constructed using the filtered correlation matrix C'. The top curve shows the predicted risk $D_{\rm p}^2$ 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 $D_{\rm r}^2$ for the same family of portfolios computed using C'_{95}. The predicted risk is now closer to the realized risk: $D_{\rm r}^2/D_{\rm p}^2 \approx 1.25$. For the same family of optimal portfolios, the dashed curve shows the realized risk computed using the original correlation matrix C_{95}: $D_{\rm r}^2/D_{\rm p}^2 \approx 1.3$.