Boston University

OpenBU   **http://open.bu.edu**

Physics   BU Open Access Articles

2012-09-10

Identifying states of a financial market

This work was made openly accessible by BU Faculty. Please share how this access benefits you. Your story matters.

Version	Published version
Citation (published version):	Michael C. Muennix, Takashi Shimada, Rudi Schaefer, Francois Leyvraz, Thomas H. Seligman, Thomas Guhr, H. Eugene Stanley. 2012. "Identifying States of a Financial Market." SCIENTIFIC REPORTS, Volume 2. https://doi.org/10.1038/srep00644

https://hdl.handle.net/2144/39959
Boston University

Identifying States of a Financial Market

Michael C. Münnix[1,2], Takashi Shimada[1,3], Rudi Schäfer[2], Francois Leyvraz[4], Thomas H. Seligman[4], Thomas Guhr[2] & H. Eugene Stanley[1]

[1]Center of Polymer Studies, Boston University, USA, [2]Faculty of Physics, University of Duisburg-Essen, Germany, [3]Department of Applied Physics, Graduate School of Engineering, The University of Tokyo, Japan, [4]Instituto de Ciencias Físicas, Universidad Nacional Autónoma de México and Centro Internacional de Ciencias, Cuernavaca, Mexico.

The understanding of complex systems has become a central issue because such systems exist in a wide range of scientific disciplines. We here focus on financial markets as an example of a complex system. In particular we analyze financial data from the S&P 500 stocks in the 19-year period 1992–2010. We propose a definition of state for a financial market and use it to identify points of drastic change in the correlation structure. These points are mapped to occurrences of financial crises. We find that a wide variety of characteristic correlation structure patterns exist in the observation time window, and that these characteristic correlation structure patterns can be classified into several typical "market states". Using this classification we recognize transitions between different market states. A similarity measure we develop thus affords means of understanding changes in states and of recognizing developments not previously seen.

T ime series are typical experimental results we have about complex systems. In the analysis of such time series, stationary situations have been extensively studied and correlations have been found to be a very powerful tool. Yet most natural processes are non-stationary. In particular, in times of crisis, accident or trouble, stationarity is lost. As examples we may think of financial markets, biological systems, reactors (both chemical and nuclear) or the weather. In non-stationary situations analysis becomes very difficult and noise is a severe problem. Following a natural urge to search for order in the system, we endeavor to define states through which systems pass and in which they remain for short times. Success in this respect would allow to get a better understanding of the system and might even lead to methods for controlling the system in more efficient ways. We here concentrate on financial markets because of the easy access we have to good data, because of our previous experience and last but not least because of the strong non-stationary effects recently seen.

Results

Using a similarity measure we were able to classify several typical market states between which the market jumps back and forth. Some of these states can easily be identified in this similarity measure. However, there are several states in which the market only stays for a short period. Thus, these states are sparsely embedded in time. With a k-means clustering analysis, we were able to identify these states and disclose a detailed dynamics of the market's state.

Our findings offer insight for constructing an "early warning system" for financial markets. By providing a simple instrument to identify similarities to previous states during an upcoming crisis, one can judge the current situation properly and be prepared to react if the crisis materializes. Certainly, an indication for a crisis is also given when the correlation structure undergoes rapid changes.

Another possible application of the similarity measure is risk management. Given the similarity measure, the portfolio manager is aware of periods in which the market behaved completely differently and thus can choose not to include them in his calculations. He can furthermore identify regions in which the market behaved similarly and refer to these regions when estimating the correlation matrix.

Our empirical study is a first step towards the identification of states in financial markets which are a prominent example of complex non-stationary systems.

Discussion

The effort to understand the dynamics in financial markets is attracting scientists from many fields[1–8]. Statistical dependencies between stocks are of particular interest, because they play a major role in the estimation of financial risk[9]. Since the market itself is subject to continuous change, the statistical dependencies also change in time. This non-stationary behavior makes an analysis very difficult[10,11]. Changes in supply and demand can even lead to a two phase behavior of the market[12,13]. Here, we use the correlation matrix to identify and classify the market state.

In particular we ask: How similar is the present market state, compared to previous states? To calculate this *similarity* we measure temporal changes in the statistical dependence between stock returns.

For stationary systems described by a (generally large) number K of time series, the Pearson correlation coefficient is extremely useful. It is defined as

$$C_{ij} \equiv \frac{\langle r_i r_j \rangle - \langle r_i \rangle \langle r_j \rangle}{\sigma_i \sigma_j}. \qquad (1)$$

Here the r_i and r_j represent the time series of which the averages $\langle \ldots \rangle$ are taken over a given time horizon T. σ_i and σ_j are their respective standard deviations. When calculating the correlation coefficients of K stocks, we obtain the $K \times K$ correlation matrix $\mathbf{C}$, which gives an insight into the statistical interdependencies of the time series under study.

It is necessary to consider data over large time horizons T so as to obtain reliable statistics. This leads to a fundamental problem that arises in the case of *non-stationary* systems: To extract useful information from empirical data we seek a correlation matrix from very recent data, in order to provide a good description of current correlation structure. This is because correlations change dynamically due to the non-stationarity of the process, making it very difficult to estimate them precisely[14–17]. However, if the length T of the time series is short, the correlation matrices $\mathbf{C}$ are noisy. On the other hand, to keep the estimation error low, T can be increased, but this leads to a correlation matrix that generally does not describe the present state very well. Various noise reduction techniques provide methods to conquer noise[18–22].

In several non-stationary systems, it is possible to obtain a large number of correlated data over time. Such systems include, but are not restricted to, financial markets (which show non-stationary behavior due to crises), biological or medical time series (such as EEG), chemical and nuclear reactors (non-stationary behavior includes, in particular, accidents) or weather data. In the following, we only consider the financial markets, since we have studied extensively some very high quality data of this system, the non-stationary features of which have been quite striking in the last

years. We propose a definition of a state which is appropriate for such systems and suggest a method of analysis which allows for a classification of possible behaviors of the system. When $T/K < 1$, which is the case we are interested in, the correlation matrix becomes singular. However, one can still make significant statistical statements, e.g., for the average correlation level whose estimation error decreases as $1/K$. In the following, we focus on correlation matrices $\mathbf{C}(t_1)$ and $\mathbf{C}(t_2)$ at different times t_1 and t_2 measured over a *short* time horizon. These have therefore a pronounced random element. We take these objects as the fundamental states of our system. We now propose, as a central element, to introduce the following concept of distance between two states. We define the similarity measure

$$\zeta(t_1, t_2) \equiv \left\langle \left| C_{ij}(t_1) - C_{ij}(t_2) \right| \right\rangle_{ij} \qquad (2)$$

to quantify the difference of the correlation structure for two points in time, where $|\ldots|$ denotes the absolute value and $\langle \ldots \rangle_{ij}$ denotes the average over all components. Note that in this case, the random component that is unavoidable in the definition of the states of the system is strongly suppressed by the average over $K^2 \gg 1$ numbers.

To apply the above general statements to a specific example, we analyze two datasets: (i) we calculate $\zeta(t_1, t_2)$ based on the daily returns of those S&P 500 stocks that remained part of the S&P during the 19-year period 1992–2010, and (ii) we study the four-year period 2007–2010 in more detail based on intraday data from the NYSE TAQ database. Since the noise increases for very high-frequency data[23–25], we extract one-hour returns for dataset (ii). For one-hour returns, we consider this market microstructure noise as reasonably weak.

However, sudden changes in drift and volatility are present on all time scales. They can result in erroneous correlation estimates. To address this problem, we employ a local normalization[26] of the return time series in dataset (i). The results of dataset (i) are presented in left panel of Fig. 1. In this figure, each point is calculated on correlation matrices over the previous two months. This new representation gives a complete overview about structural changes of this financial market of the past 19 years in a single figure. It allows to compare the similarity of the market states at different times. To make this
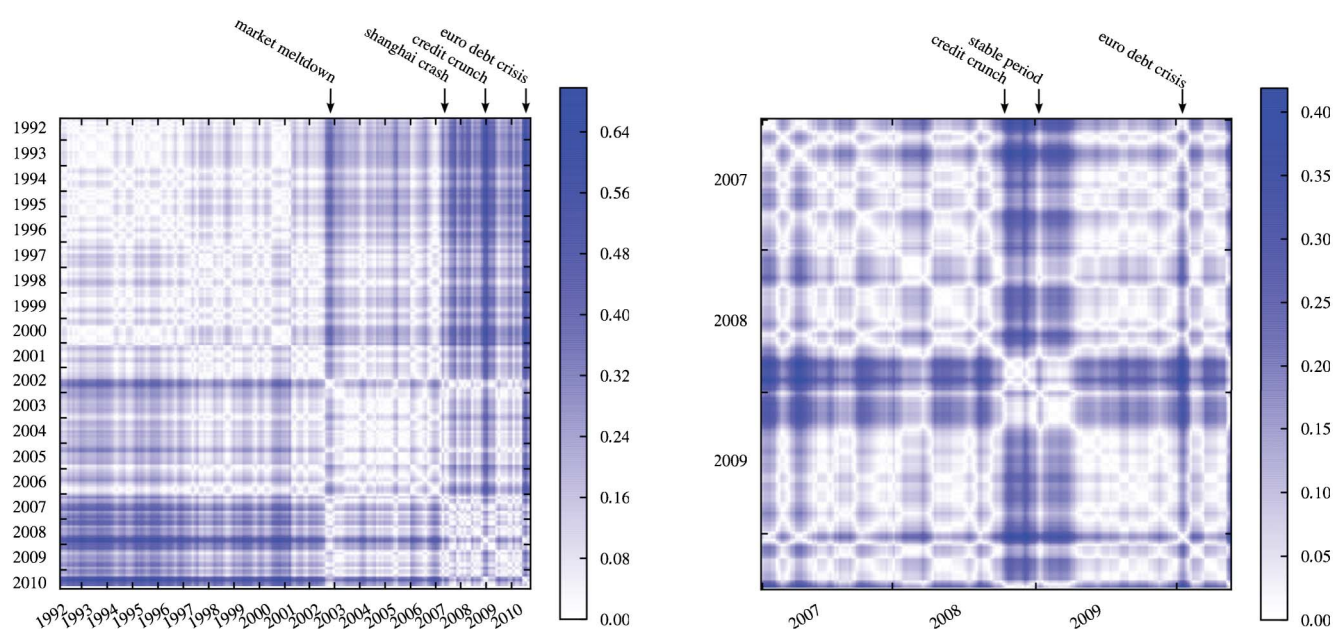

Figure 1 | Financial crisis are accompanied by drastic changes in the correlation structure, indicated by blue shaded areas. The market similarity ζ in the left panel is based on daily data. The right panel is a more detailed study of the 2007–2010 period, including the "credit crunch." The area of the right panel is a magnification of the lower right square in the left panel.
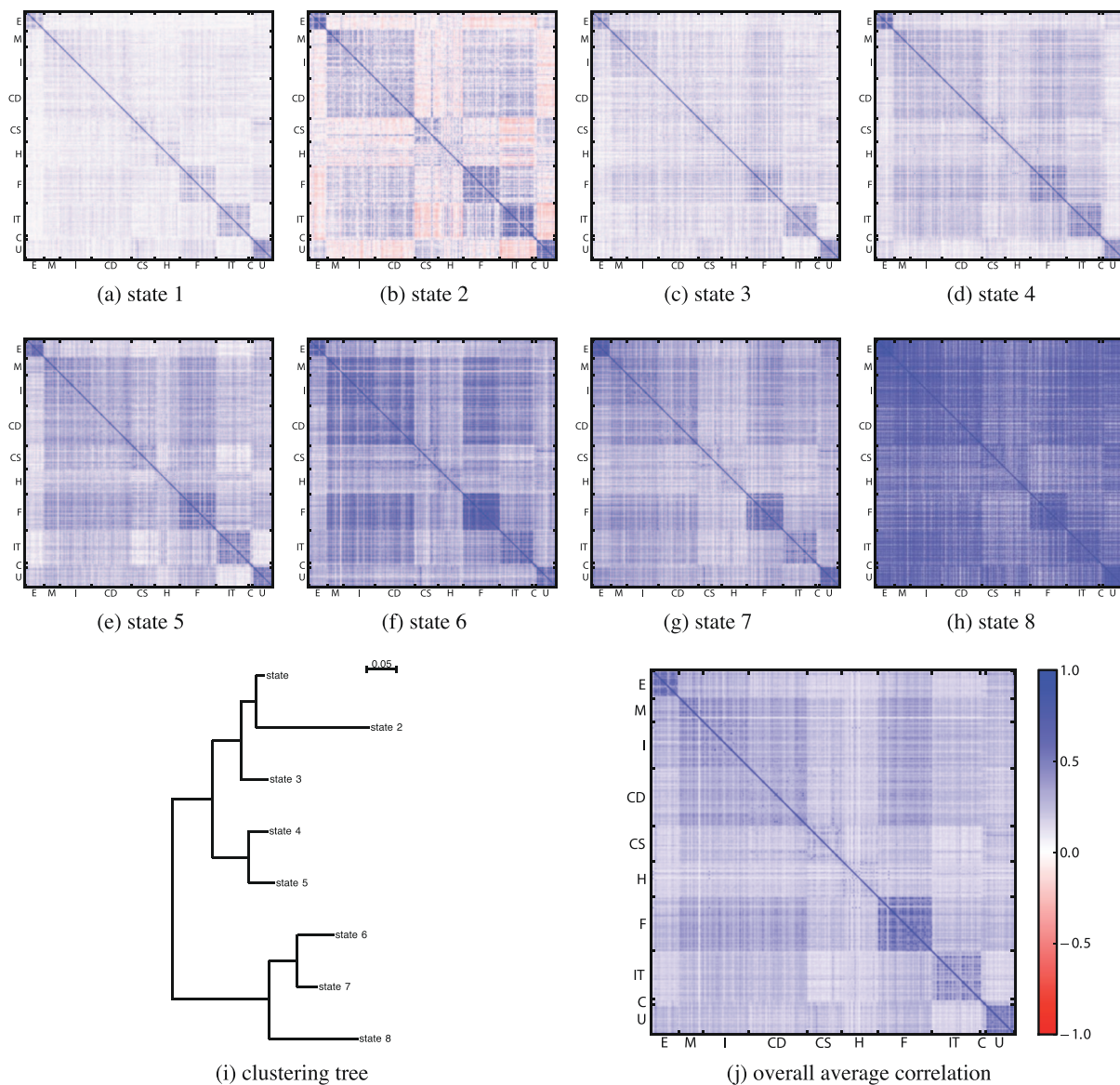
Figure 2 | **The correlation between different industry branches as well as the intra-branch correlation characterize the different market states (a-h).** The inter-branch correlation is represented by the off-diagonal blocks, and the intra-branch correlation is represented by the blocks on the diagonal. Legend: E: Energy, M: Materials, I: Industrials, CD: Consumer Discretionary, CS: Consumer Staples, H: Health Care, F: Financials, IT: Information Technology, C: Communication, U: Utilities. (i) Similarity tree structure of the 8 market states. (j) Illustration of the overall average correlation matrix.

procedure concrete, consider the following example. Pick a point on the diagonal of this panel and designate it as "now". From this point the similarity to previous times can be found on the vertical line above this point, or the horizontal line to the left of this point. Light shading denotes similar market states and dark shading denotes dissimilar states. We can furthermore identify times of financial crises with dark shaded areas. This indicates that the correlation structure completely changes during a crisis. There are also similarities between crises, as between the "credit crunch" that induced the 2008–2009 financial crisis and the "market meltdown", the burst of the "dot-com bubble" in 2002. A further example is the overall rise in correlation level in the beginning of 2007. This event can be mapped to drastic events on the Shanghai stock exchange[27].

Using dataset (ii) we are able to obtain a more detailed insight into recent market changes, as shown in the right panel of Fig. 1. This area is represented by the lower right square of the left panel. Using intra-day data we calculate the correlation matrices on shorter time scales. We choose a time horizon of one week because it provides insight into changes in the correlation structure on a much finer time scale.

This enables us to identify a short sub-period within the 2008–2009 crisis (in the beginning of 2009) during which the market temporarily stabilizes before it returns to the crisis state. While the correlation structure during the crisis displays an overall high correlation level, the correlation structure of the stable period is similar to the period before the crisis, one of the typical states in a calm period, which is identified from daily data in dataset (i). This phenomenon might be related to the market's reaction to news about the progress in rescuing the American International Group (A.I.G.)[28]. The correlation structure of the stable period and the 2008–2009 crisis is shown in Supplementary Fig. S2 online.

The evolutionary structure presented in Fig. 1 illustrates that the correlation matrix sometimes maintains its structure for a long time (bright regions), sometimes changes abruptly (sharp blue stripes), and sometimes returns to a structure resembling a structure the market has experienced before (white stripes). This suggests that the market might move among several typical market states. To extract such typical market states, we perform a clustering analysis in the results of dataset (i). From our clustering analysis (see section
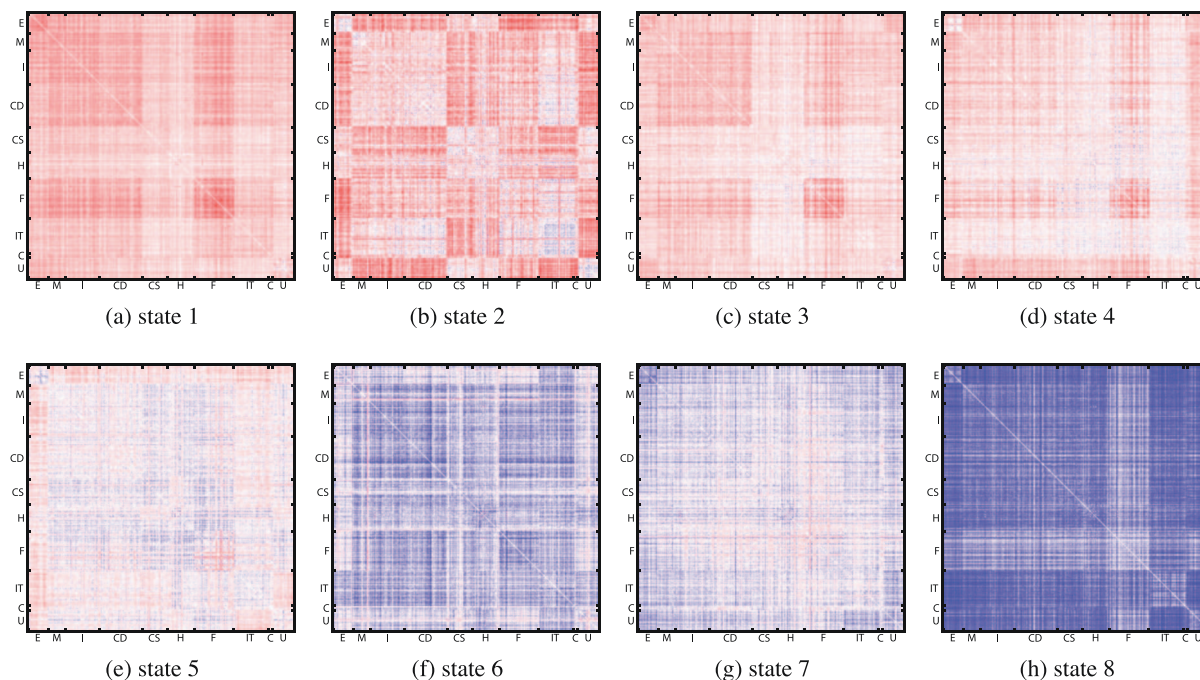
Figure 3 | (a)–(h): Difference of the states' correlation matrices to the average correlation matrix. The color-scale is identical to Fig. 2.

Methods), we find that there are "hidden" states sparsely embedded in time, in addition to regimes that dominate the market during a continuous period and are easily found by eye. The clustering analysis separates the set of all correlation matrices, which are measured on short disjoint subintervals of the observation period, into distinct clusters. The cluster centers then correspond to distinct correlation structures which we identify as market states. We note that this definition of market state depends on the length of the full observation period to some extend. Enlarging the observation period might join previously distinct clusters, reducing it might further divide a cluster.

For the clustering analysis, we use disjunct two-month time windows ending at the respective dates. Because of the window length, some financial crashes cannot be resolved. Our aim is rather to identify the evolution of the market, which is, in some cases, induced by financial crisis. We can confirm in Fig. 2 that the typical states obtained from the clustering analysis indeed correspond to different characteristic correlation structures. To visualize the characteristic structures of each state, we calculate its average correlation matrix and sort the companies according to their industry branch, as defined by the Global Industry Classification Standard (GICS)[29]. In the resulting matrices, the industry branches correspond to the blocks on the diagonal. The correlation between two branches are given by the off-diagonal blocks. The results are illustrated in Fig. 2. We can confirm that the typical states

obtained from the clustering analysis indeed correspond to different characteristic correlation structures.

Our analysis also offers insight into market structure dynamics. Figure 4 shows the temporal behavior of the market state. The market sometimes remains for a long time in the same state, and sometimes stays only for a short time. The typical duration depends upon the state: Some states (e.g., state 1 and state 2) appear in clusters in time while other states appear more sparsely in time (e.g., state 4). There seems to exist a global trend on a long time scale, although the market state is switching back and forth between states.

In Fig. 2, we can see differences between the states in the correlation between branches as well as in the correlation within a branch. The correlation within the energy, information technology, and utilities branches is very strong in all states. State 1 shows an overall weak correlation, while states 3 and 4 feature in addition a strong correlation of the finance branch to other branches. State 2 shows very unusual behavior: In the period of the dot-com bubble, many branches are anti-correlated with one another. In states 5, 6 and 7, the overall correlation level rises, although certain branches, such as energy, consumer staples, and utilities, are either strongly or weakly correlated with other branches. While some distinct characteristics of states can be easily identified, some states look quite similar by eye. Their unique properties can be illustrated when subtracting the market's overall correlation structure, as illustrated in Fig. 3. For example, state 3 and 4 look very similar in Fig. 2. However, Fig. 3
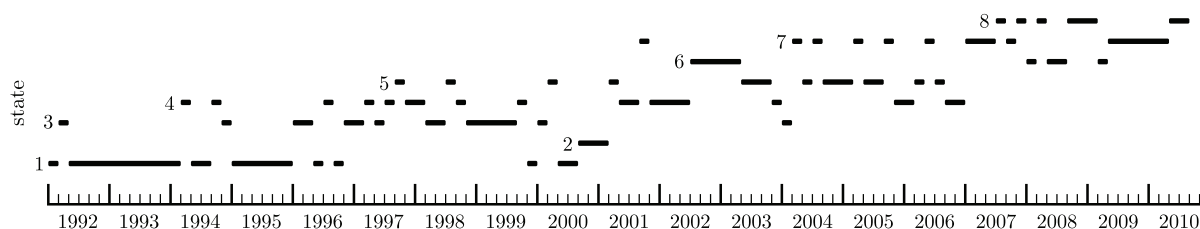

Figure 4 | Temporal evolution of the market state. The horizontal axis represents the observation time and the vertical axis denotes the market state obtained from top-down clustering. The market state sometimes remains in the same state for a long time, and sometimes only for a short time. It also can return to a state that it has previously visited. Some states (e.g., state 1 and state 2) appear to cluster in time, while other states appear more sparsely and intermittently in time (e.g., state 4).
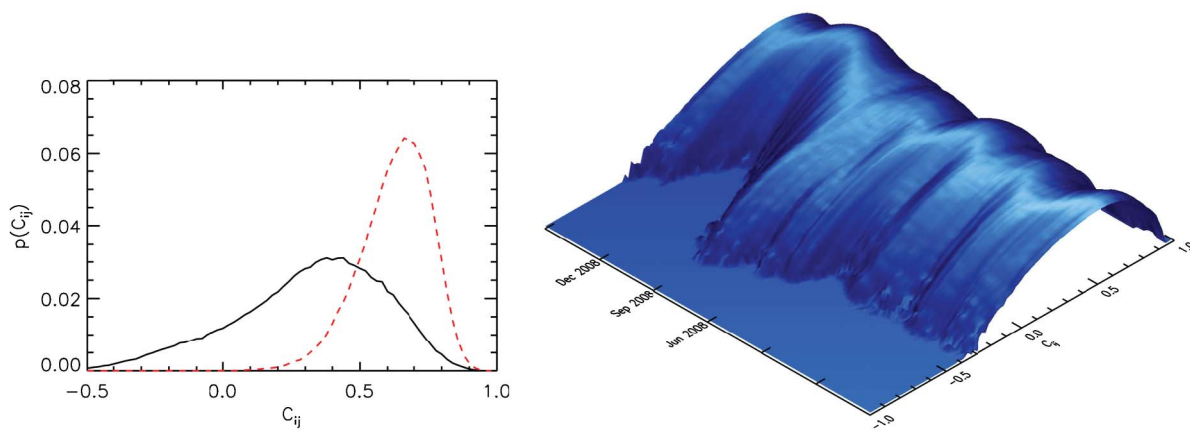
4

Figure 5 | Footprint of the state transition in the 2008 crisis by histograms of the correlation coefficients $C_{ij}(t)$. (a) Surface plot for the time period September 2007 to March 2009. We use a logarithmic scale to show the bimodal structure more clearly. (b) Histograms for September 2008 (black solid line) and December 2008 (red dashed line).

unveils that the correlation within the Energy sector (E) is completely diverse.

We observe that the energy branch (E) is either strongly correlated to the rest of the market, weakly correlated, or even anti-correlated. To analyze this behaviour we study the histogram of the correlation coefficients $C_{ij}(t)$. We present the results in Fig. 5. In the months leading up to the credit crunch in October 2008, we observe a bimodal structure in the histogram (see also Ref. [30]). It corresponds to the time period when the Energy branch shows a strong anti-correlation with other branches. The bimodality suggests that a subset of stocks – in this case, predominantly the Energy stocks – decouples from the rest of the market. During the crash, the histogram shows a very narrow distribution around large values of the correlation coefficients, which corresponds to state 8 in Fig. 2, where the branch structure is lost almost completely in an overall strongly correlated market.

Methods

Construction of stock returns. Let S be the price of a specific stock and Δt the interval on which the return is calculated. For our study, we chose the arithmetic return, defined as

$$r(t) \equiv \frac{S(t+\Delta t) - S(t)}{S(t)}. \qquad (3)$$

For dataset (i), we chose Δt to be 1 day and calculate the stock returns of each day. For dataset (ii), we chose Δt as 1 hour. Furthermore, we obtain this 1-hour return for every minute of a trading day between 10:45 am and 2:45 pm. We obtained the daily data of dataset (i) from *finance.yahoo.com.* The intraday data are obtained from the New York Stock Exchange's TAQ database.

Local normalization. Sudden changes in drift and volatility can result in erroneous correlation estimates. To address this problem, we employ a local normalization method[26]. For each return $r(t)$ we subtract the local mean and divide by the local standard deviation,

$$\tilde{r}(t) \equiv \frac{r(t) - \langle r(t) \rangle_n}{\sqrt{\langle r^2(t) \rangle_n - \langle r(t) \rangle_n^2}}. \qquad (4)$$

The local average $\langle \ldots \rangle_n$ runs over the n most recent sampling points. For daily data, $n = 13$ yields nearly normal distributed time series, as discussed in Ref. [26].

Outline of top-down clustering. Our clustering analysis is based on a top-down scheme: All m correlation matrices are initially regarded as a single cluster and then divided into sub-clusters by a procedure based on the k-means algorithm[31–33]. The process can be described as follows:

1. Choose two initial cluster centers from all matrices. Label all other matrices by the more similar cluster center in terms of $\xi^{(L)}$.

   (a) Recast two new sub-cluster centers to the "center of mass" of the matrices in the sub-cluster.
   (b) Re-label all matrices to the new sub-cluster center.
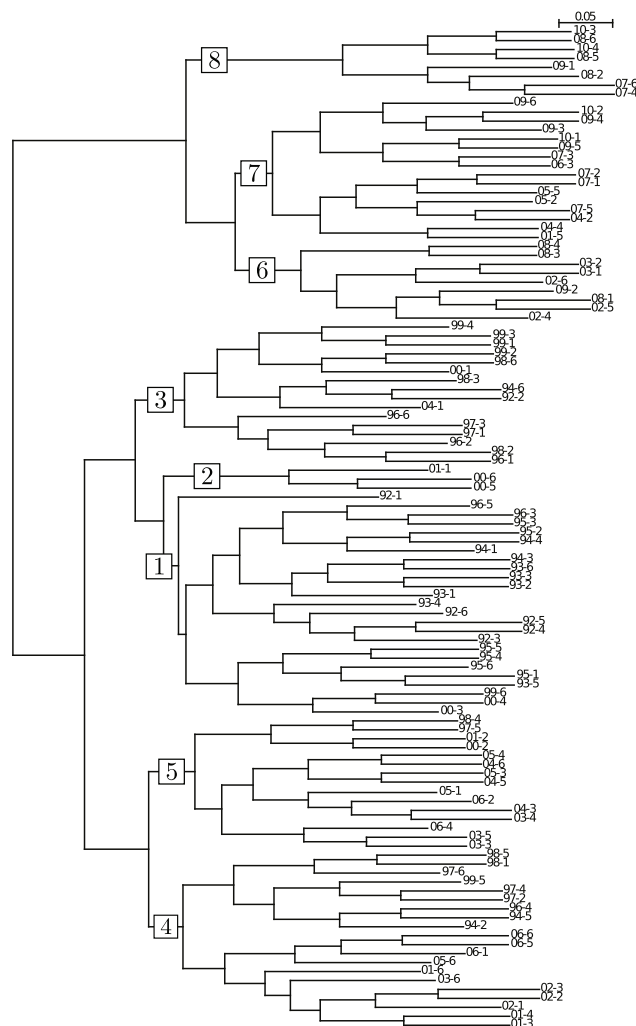   (c) Repeat this process until there is no change in labeling.

Figure 6 | The entire tree of the clustering analysis presented here for the threshold 0: No termination of the division process takes place until all the correlation matrices are identified as different components. The large bold numbers represent the market states each of which consists of the matrices in the sub-trees below. Each right end of the tree corresponds to each 2-month term (year-term). Terms 1, 2, …, 6 correspond to January to February, March to April, …, November to December, respectively. The length of each branch represents the distance from the center of the subcluster to the center of the original cluster before the last dual division.

5

2. Take the best division out of all possible $m(m-1)/2$ initial choices, which gives the least average $(\xi^{(L)})^2$ to the cluster centers.

We stop this division process when the average distance from each cluster center to its members becomes smaller than a certain threshold. To identify the typical market states presented in the manuscript, we chose the threshold at 0.1465 as it represents (approximately) the best ratio of the distances between clusters and their intrinsic radius. One can obtain finer structures by choosing smaller threshold values, ultimately until all the matrices are identified as different components. The complete results of the clustering analysis is illustrated in Fig. 6.

A possible alternative method for obtaining the states may be choosing a certain number of cluster centers from the beginning. However, of besides being more computationally intense, we believe that predefining the number of states would be a significant disadvantage compared to our dynamic approach.

Alternative measure: Difference of largest eigenvalue of correlation matrices. A similar result can be archived using a different approach. The largest eigenvalue λ_{max} of the correlation matrix $\mathbf{C}$ describes the collective motion of all stocks. We can also define the similarity measure by the distance of these eigenvalues,

$$\xi_{alt}(t_1,t_2) \equiv |\lambda_{max}(\mathbf{C}(t_1)) - \lambda_{max}(\mathbf{C}(t_2))|. \qquad (5)$$

The advantage of this technique is that the noise in the correlation matrix only contributes to small eigenvalues (See Refs. [18] and [19]). Thus, by only taking into account the largest one, we filter out the noise. However, this approach also presumes that the corresponding eigenvector does not change. Our results indicate that the largest eigenvalue almost remains constant, but this might not always be the case. Especially in financial crises. An alternative distance measure which takes into account the changes in the eigenvectors is considered in Refs. [34, 35]. The obtained similarity measure is shown in Supplementary Fig. S1 online.

1. Voit, J. *The Statistical Mechanics of Financial Markets* (Springer, Heidelberg, 2001).
2. Mantegna, R. N. & Stanley, H. E. Stock market dynamics and turbulence: parallel analysis of fluctuation phenomena. *Physica A* **239**, 255–266 (1997).
3. Borghesi, C., Marsili, M. & Miccichè, S. Emergence of time-horizon invariant correlation structure in financial returns by subtraction of the market mode. *Physical Review E* **76**, 026104 (2007).
4. Cooley, T. F. & Quadrini, V. Financial markets and firm dynamics. *The American Economic Review* **91**, 1286–1310 (2001).
5. Pelletier, D. Regime switching for dynamic correlations. *Journal of Econometrics* **131**, 445–473 (2006).
6. Xu, X. E., Chen, P. & Wu, C. Time and dynamic volume-volatility relation. *Journal of Banking & Finance* **30**, 1535–1558 (2006).
7. King, M. & Wadhwani, S. Transmission of volatility between stock markets. *The Review of Financial Studies* **3**, 5–33 (1990).
8. Lee, S. The stability of the co-movements between real estate returns in the uk. *Journal of Property Investment & Finance* **24**, 434–442 (2006).
9. Bouchaud, J. & Potters, M. *Theory of Financial Risks* (Cambridge University Press, Cambridge, 2000).
10. Eckmann, J.-P., Kamphorst, S. O. & Ruelle, D. Recurrence plots of dynamical systems. *EPL (Europhysics Letters)* **4**, 973 (1987).
11. Casdagli, M. C. Recurrence plots revisited. *Physica D: Nonlinear Phenomena* **108**, 12–44 (1997).
12. Plerou, V., Gopikrishnan, P. & Stanley, E. Econophysics: Two-phase behaviour of financial markets. *Nature* **421** (2003).
13. Plerou, V., Gopikrishnan, P. & Stanley, H. E. Two phase behaviour and the distribution of volume. *Quantitative Finance* **5**, 519–521 (2005).
14. Rosenow, B., Gopikrishnan, P., Plerou, V. & Stanley, H. E. Dynamics of cross-correlations in the stock market. *Physica A* **324**, 241–246 (2003).
15. Tastan, H. Estimating time-varying conditional correlations between stock and foreign exchange markets. *Physica A* **360**, 445–458 (2006).
16. Drozdz, S., Kwapien, J., Grmmer, F., Ruf, F. & Speth, J. Quantifying the dynamics of financial correlations. *Physica A* **299**, 144–153 (2001).
17. Schäfer, R., Nilsson, N. & Guhr, T. Power mapping with dynamical adjustment for improved portfolio optimization. *Quantitative Finance* (2009).
18. Laloux, L., Cizeau, P., Bouchaud, J.-P. & Potters, M. Noise dressing of financial correlation matrices. *Physical Review Letters* **83**, 1467–1470 (1999).
19. Plerou, V. *et al.* Random matrix approach to cross correlations in financial data. *Physical Review E* **65**, 066126 (2002).
20. Guhr, T. & Kälber, B. A new method to estimate the noise in financial correlation matrices. *Journal of Physics A: Mathematical and General* **36**, 3009–3032 (2003).
21. Schäfer, J. & Strimmer, K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology* **4** (2005).
22. Ledoit, O. & Wolf, M. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance* **10**, 603–621 (2003).
23. Epps, T. W. Comovements in stock prices in the very short run. *Journal of the American Statistical Association* **74**, 291–298 (1979).
24. Münnix, M. C., Schäfer, R. & Guhr, T. Compensating Asynchrony Effects in the Calculation of Financial Correlations. *Physica A* **389**, 767–779 (2010).
25. Münnix, M. C., Schäfer, R. & Guhr, T. Impact of the Tick-size on Financial Returns and Correlations. *Physica A* **389**, 4828–4843 (2010).
26. Schäfer, R. & Guhr, T. Local normalization: Uncovering correlations in non-stationary financial time series. *Physica A* **389**, 3856–3865 (2010).
27. What The Market Is Telling Us. *Bloomberg Businessweek* (2007). Cover story.
28. A.I.G. Sells $39.3 Billion in Assets to N.Y. Fed's Fund'. *The New York Times* (2008).
29. http://www.standardandpoors.com/indices/gics/en/us. accessed on 16 Aug 2012.
30. Wyart, M. & Bouchaud, J.-P. Self-referential behaviour, overreaction and conventions in financial markets. *Journal of Economic Behavior & Organization* **63**, 1–24 (2007).
31. MacQueen, J. B. Some methods for classification and analysis of multivariate observations. In Cam, L. M. L. & Neyman, J. (eds.) *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability,* vol. 1, 281–297 (University of California Press, 1967).
32. Faber, V. Clustering and the continuous k-means algorithm. *Los Alamos Science* **22**, 138–144 (1994).
33. Tanaka, N. K., Awasaki, T., Shimada, T. & Ito, K. Integration of chemosensory pathways in the drosophila second-order olfactory centers. *Current biology* **14**, 449–457 (2004).
34. Reigneron, P.-A., Allez, R. & Bouchaud, J.-P. Principal regression analysis and the index leverage effect. *Physica A* **390**, 3026–3035 (2011).
35. Allez, R. & Bouchaud, J.-P. Eigenvector dynamics: theory and some applications Preprint: arXiv:1108.4258. (2011).

Acknowledgments

Author contribution

MCM, TS and RS performed the calculations and analyzed the results. MCM and RS prepared the figures. All authors designed the research and wrote and reviewed the paper.

Additional information