New Journal of Physics
The open access journal at the forefront of physics

Deutsche Physikalische Gesellschaft DPG
IOP Institute of Physics

Published in partnership with: Deutsche Physikalische Gesellschaft and the Institute of Physics

OPEN ACCESS

RECEIVED
11 June 2018

REVISED
11 September 2018

ACCEPTED FOR PUBLICATION
12 October 2018

PUBLISHED
31 October 2018

PAPER

Identifying long-term precursors of financial market crashes using correlation patterns

Hirdesh K Pharasi[1,5], **Kiran Sharma**[2], **Rakesh Chatterjee**[1,3], **Anirban Chakraborti**[2,5], **Francois Leyvraz**[1,4] **and Thomas H Seligman**[1,4]

[1] Instituto de Ciencias Físicas, Universidad Nacional Autónoma de México, Cuernavaca-62210, México
[2] School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi-110067, India
[3] School of Mechanical Engineering and Sackler Center for Computational Molecular and Materials Science, Tel Aviv University, Tel Aviv-6997801, Israel
[4] Centro Internacional de Ciencias, Cuernavaca-62210, México
[5] Author to whom any correspondence should be addressed.

E-mail: hirdeshpharasi@gmail.com **and** anirban@jnu.ac.in

Keywords: market crash, market state, power mapping method, multidimensional scaling, return cross-correlations

Supplementary material for this article is available online

Abstract

The study of the critical dynamics in complex systems is always interesting yet challenging. Here, we choose financial markets as an example of a complex system, and do comparative analyses of two stock markets—the S&P 500 (USA) and Nikkei 225 (JPN). Our analyses are based on the evolution of cross-correlation structure patterns of short-time epochs for a 32 year period (1985–2016). We identify 'market states' as clusters of similar correlation structures, which occur more frequently than by pure chance (randomness). The dynamical transitions between the correlation structures reflect the evolution of the market states. Power mapping method from the random matrix theory is used to suppress the noise on correlation patterns, and an adaptation of the intra-cluster distance method is used to obtain the 'optimum' number of market states. We find that the S&P 500 is characterized by four market states and Nikkei 225 by five. We further analyze the co-occurrence of paired market states; the probability of remaining in the same state is much higher than the transition to a different state. The transitions to other states mainly occur among the immediately adjacent states, with a few rare intermittent transitions to the remote states. The state adjacent to the critical state (market crash) may serve as an indicator or a 'precursor' for the critical state and this novel method of identifying the long-term precursors may be helpful for constructing the early warning system in financial markets, as well as in other complex systems.

1. Introduction

A financial market is a highly complex and continuously evolving system [1–3]. To understand the statistical behavior of the financial market and its constituent sectors [4–9], researchers focused their attention on the information of co-movements and correlations among the stocks of the market. It is well known that the mean correlation among the stocks assumes much higher values during market crashes than in normal business periods [10]. Similarly, certain correlation structures seem to occur more frequently than by pure chance (randomness), specially when markets approach a critical period or crash [11, 12]. However, to identify such similar (clusters) correlation patterns, referred as 'market states', as was previously attempted by Munnix *et al* [13, 14], is rather challenging due to many factors. The first factor is that financial time series is non-stationary; second factor is that there is always noise present in the correlations computed over finite length time series data [15], and it is essential to suppress the corresponding noise in correlation matrices to reveal the actual correlations. To tackle the first factor of non-stationarity, we work with short time series so that the number of time steps over which we compute the correlations can be considered as reasonably stationary. However, with

short time series the correlation matrices become highly singular [16–18]. To tackle the second factor of noise reduction, various techniques [19, 20] are available. Here, we shall use a recent and efficient one, namely the power map method [19, 21, 22], for noise reduction as well as breaking the degeneracy in the eigenvalues so that the correlation matrices are no longer singular. Furthermore, the problem of finding similar clusters (groups) of the correlation patterns is a daunting task by itself. To go beyond the simple quantification of financial market states in terms of the average correlation, clustering techniques seem promising as does the study of eigenvalues of the correlation matrix of the corresponding time series [15]. In the research of clustering, the *k*-means method has had some success for top-to-down clustering, but it suffers from one major drawback: the number of clusters is ad hoc. Earlier, Munnix *et al* [13] had provided a scheme where all the correlation matrices at different epochs were initially regarded as a single cluster and then divided into sub-clusters by a procedure based on the *k*-means algorithm. They stopped the division process when the average distance from each cluster center to its members became smaller than a certain threshold. Based on the top-to-down hierarchical clustering method and the threshold at 0.1465, which represented the best ratio of the distances between clusters and their intrinsic radii, Munnix *et al* had determined the number of markets states for USA to be eight. In the present paper, for determining the 'optimal' number of clusters, we use multidimensional scaling (MDS) technique [23] with two/three-dimensional representations, which are comparatively easier for visualization and studying time evolution. So, using MDS map, we apply *k*-means clustering to divide the clusters of similar correlation patterns into *k* groups. We propose a new way, based on the variance of cluster radii, for estimating the number of clusters *k*, which is fairly robust and stable. We thus have a considerable degree of confidence in determining the 'optimal' number of market states identified by the new prescription. For our research, we have used adjusted closure price data from Yahoo finance [24] for the S&P 500 (USA) and Nikkei 225 (JPN) stock exchanges, for the 32 year period (1985–2016). The stock list has been filtered such that we have only stocks which were included in the market index for the entire period of 32 years. Among others, our main finding is that there exist four market states in USA and five in JPN. We then study the dynamical transitions between the market states, in a probabilistic manner; we also analyze the co-occurrence of paired market states and find that the probability of remaining in the same state is much higher than jumping to another state. The transitions mainly occur among adjacent states, with a few rare intermittent transitions to the remote states. The state adjacent to the critical state may indicate a 'precursor' to the critical state (market crash) and this new method of identifying the long-term precursors may be helpful for constructing the early warning system in financial markets, and in other complex systems.

The paper is organized as follows: we present briefly the methodology and the data description. Then we present the main part of data analyses along with the above mentioned findings; additional details can be found in the supplementary information is available online at stacks.iop.org/NJP/20/103041/mmedia. Finally, we present summary and concluding remarks.

2. Data description, methodology and results

2.1. Data description

We have used the database of Yahoo finance [24], for the time series of adjusted closure price for two countries: S&P 500 (USA) index and Nikkei 225 (JPN) index, for the period 02-01-1985 to 30-12-2016, and for the corresponding stocks as follows:

- USA—02-Jan-1985 to 30-Dec-2016 ($T = 8068$ d); Number of stocks $N = 194$;
- JPN—04-Jan-1985 to 30-Dec-2016 ($T = 7998$ d); Number of stocks $N = 165$,

where we have included the stocks which are present in the indices for the entire duration. The sectoral abbreviations are given in table 1.

The list of stocks (along with the sectors) for the two markets are given in the tables S1 and S2 in supplementary information.

2.2. Cross-correlation matrix and power mapping method

We present a study of time evolution of the cross-correlation structures of return time series for N stocks, and determination of the optimal number of market states (correlation patterns that exist more frequently then by pure chance or randomness); also, the dynamical evolution of the market states over different epochs. The daily return time series is constructed as $r_k(t) = \ln P_k(t) - \ln P_k(t-1)$, where $P_k(t)$ is the adjusted closing price of the kth stock at time t (trading day). Then, the cross-correlation matrix is constructed using equal-time Pearson cross-correlation coefficients, $C_{ij}(\tau) = (\langle r_i r_j \rangle - \langle r_i \rangle \langle r_j \rangle)/\sigma_i \sigma_j$, where $i, j = 1, ..., N$, τ indicates the end date of

Table 1. Abbreviations of different sectors for S&P 500 and Nikkei 225 markets.

Labels	Sectors	Labels	Sectors
CD	Consumer discretionary	**ID**	Industrials
CS	Consumer staples	**IT**	Information technology
CP	Capital goods	**MT**	Materials
CN	Consumer goods	**PR**	Pharmaceuticals
EG	Energy	**TC**	Technology
FN	Financials	**UT**	Utilities
HC	Health care		

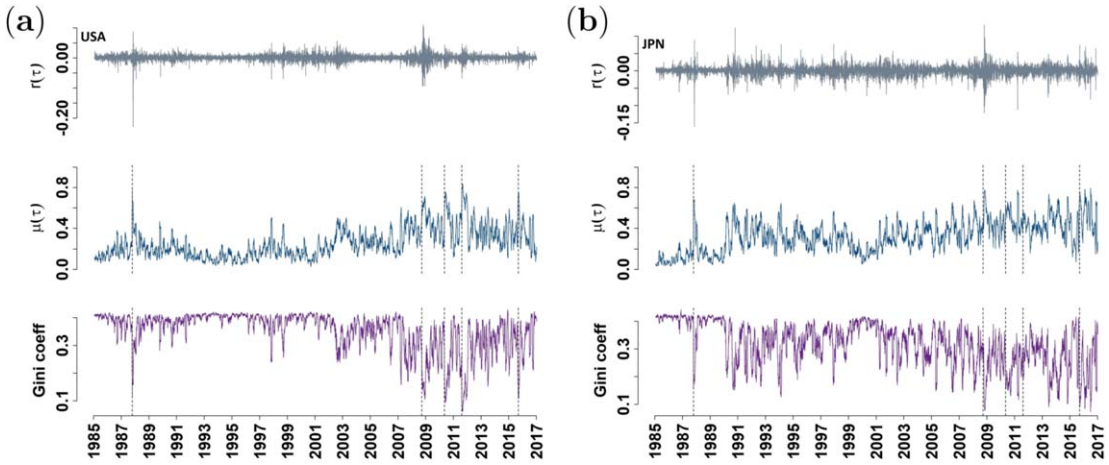

Figure 1. Results of market evolution for (a) USA and (b) JPN, respectively. The top row shows the returns of the respective market indices. The middle row shows the mean market correlation (averaged over all the cross-correlation coefficients) of the respective markets. The bottom row shows the inequality in the distribution of the cross-correlation coefficients, as characterized by the Gini coefficient. Evidently, whenever there is a market crash, the mean market correlation becomes high and the Gini coefficient becomes low, indicating that all the stocks behave similarly.

the epoch of size M d, and $\langle ... \rangle$ as well as the standard deviations are computed over that epoch. Here, we computed daily return cross-correlation matrix $\boldsymbol{C}(\tau)$ computed over the short epoch of $M = 20$ d, for (a) USA with $N = 194$ stocks of S&P 500 for a return series of $T = 8060$ d, and (b) JPN with $N = 165$ stocks of Nikkei 225 for $T = 7990$ d, during the calendar period 1985–2016. We use epochs of 20 d to obtain a balance between choosing short epochs for detecting changes and long ones for reducing fluctuations. In figure 1, we show the time evolution of the return of the market index, $r(\tau)$, along with the mean market correlation (average of all the elements of the cross-correlation matrix), $\mu(\tau)$, and the Gini coefficient that characterizes the variation in the distribution of the correlation coefficients. Evidently, whenever there is a market crash (fall in the $r(\tau)$), the mean market correlation $\mu(\tau)$ rises a lot, and the Gini coefficient falls drastically, indicating that market is extremely correlated and most of the stocks behave similarly (see [10]). Since the assumption of stationarity manifestly fails for longer return time series, it is often useful to break the long time series of length T, into shorter epochs of size M (such that $T/M = n$). The assumption of stationarity improves for the shorter epochs used. As mentioned in the introduction, we use the power map technique [19, 21, 22] to suppress the noise present in the correlation structure of short time series. In this method, a nonlinear distortion is given to each cross-correlation coefficient within an epoch by: $C_{ij} \rightarrow (\text{sign } C_{ij})|C_{ij}|^{1+\epsilon}$, where ϵ is the noise-suppression parameter. This also gives rise to an 'emerging spectrum' of eigenvalues, arising from the breaking of the degeneracy of the zero eigenvalues (see [15, 22] for recent reviews).

2.3. Noise-suppression in a short time cross-correlation matrix

First, we study the effect of the noise-suppression parameter ϵ on the cross-correlation matrix and its eigenvalue spectrum within an epoch. The cross-correlation structure can be visualized easily through a two/three-dimensional map of coordinates generated through a MDS algorithm. The MDS is a tool of nonlinear dimensional reduction to visualize the similarity of the data set in a D-dimensional space. Each object is assigned to a coordinate space in D-dimensional space keeping the between-object distance preserved, as close as

possible. The choice of $D = 2$ or $D = 3$ is for optimizing the object location to two/three-dimensional scatter plots or maps. As an input to the MDS algorithm, we provide the distance matrix [25], generated from the correlation matrix, using the nonlinear transformation:

$$d_{ij} = \sqrt{2(1 - C_{ij})}.$$

The effect of the variation of the parameter ϵ on noise reduction and determining the optimal number of market states, can thus be better captured through the MDS. The question is *what should be the ideal choice of the noise-suppression parameter* ϵ? A very small value of ϵ, say $\epsilon = 0.01$, surely breaks the degeneracy of eigenvalues (giving rise to an 'emerging spectrum' with interesting properties [10]) but does not contribute much to noise-suppression. On the other hand, a large value, say $\epsilon = 0.5$, suppresses the noise in the correlation pattern and thus improves clustering; however, the emerging spectrum tends to approach the original spectrum or even overlap with it. Furthermore, it also distorts the original spectrum to some extent. Yet we know, that the information is optimized in some sense (see [22] section 2), and the basic information of the largest eigenvalues determining the market and the market sectors are not significantly distorted. We need the noise suppression not only to get good clustering, as in [13], but beyond that, it is the dependence of the intra-cluster distance on the noise-suppression parameter, which proves crucial to the determination of the optimal state number. Hence, we use $\epsilon = 0.6$ and this choice of a high value is based on the robustness and finding distinct clusters of stocks using MDS. The effect can be seen through the supplementary figures S2 and S3. Further, our main aim is to find the optimal number of market states, based on correlation structures which are similar and appear more frequently. Hence, we formulate a similarity measure, ζ (to be defined later) between different cross-correlation matrices at different epochs τ, and then find similar groups of correlation matrices across different epochs. We find that with $\epsilon = 0.6$, the noise-suppressed cross-correlation structures can be grouped well into similar clusters (as we will describe later). While we find that the number of market states is not very sensitive to the noise-suppression parameter but the dependence will be useful later. A higher value of ϵ lowers the mean of the cross-correlation coefficients, μ (see supplementary figure S1) and the maximum eigenvalue $\lambda_{\max}$ of the cross-correlation matrix.

Figure 2 shows the effect of noise-suppression using power mapping method [10, 16, 19, 26] on the short time cross-correlation matrix . Figure 2(a) shows a correlation matrix computed for the short epoch $M = 20$ d for USA with $N = 194$ stocks of S&P 500 ending on 30/11/2001 (arbitrarily chosen date). The corresponding MDS map of the correlation matrix is shown in figure 2(d). For any short time series $M < N$, the highly singular correlation matrices will have $N - M + 1$ degenerate eigenvalues at zero. Hence, in our case the eigenvalue spectrum consists of 175 eigenvalues at zero, followed by 19 distinct positive eigenvalue. The nonlinear power mapping method removes the degeneracy of eigenvalues at zero, leading to an emerging spectrum [10, 15]. Figure 2(b) shows the correlation pattern for $\epsilon = 0.01$. The effect of the small distortion on the corresponding MDS map is shown in figure 2(e). The effect is less visible on MDS map for small distortion. Next, we use a high value of noise-suppression parameter $\epsilon = 0.6$ to reduce considerably the noise of the correlation matrix (shown in figure 2(c)). The effect of $\epsilon = 0.6$ on corresponding MDS map is strong, as shown in figure 2(f). Note that the clusters of stocks in the MDS maps are distinct and denser as compare to low noise-suppression ($\epsilon = 0.01$) or without noise-suppression ($\epsilon = 0$).

2.4. Noise-suppression in a similarity matrix among correlation matrices over different epochs

The noise-suppressed cross-correlation structures of return matrices $\boldsymbol{C}(\tau)$ across different times $\tau = 1, \ldots, n$, can be compared based on their similarities. If there are two correlation matrices $C(\tau_1)$ and $C(\tau_2)$ at different epochs τ_1 and τ_2, each computed over a short epoch of M d, then to quantify the similarity between the correlation structures, the similarity measure is computed as: $\zeta(\tau_1, \tau_2) \equiv \langle |C_{ij}(\tau_1) - C_{ij}(\tau_2)| \rangle$, where $|...|$ denotes the absolute value and $\langle ... \rangle$ denotes the average over all matrix elements $\{ij\}$ [13]. We then use the MDS map to visualize the information contained in $n \times n$ similarity matrix, where each element is $\zeta(\tau_p, \tau_q)$, where p, $q = 1, \ldots n$.

Interestingly, the noise-suppression applied to individual correlation matrices in short epochs, has a dramatic effect in the similarity matrix too. Figure 3 shows the effect of noise-suppression on the similarity matrix [13] and the corresponding MDS map. Each correlation matrix is computed with $N = 194$ stocks of S&P 500; hence, for the time series of length $T = 8060$ d during the period 1985–2016, there are $n = 805$ correlation matrices constructed from short epochs of $M = 20$ d and shifts of $\Delta\tau = 10$ d (50% overlapping epochs). Similarly, we have $N = 165$ stocks of Nikkei 225; the time series of length $T = 7990$ d in the same period yield $n = 798$ correlation matrices. The sharp changes in the structural patterns of the similarity matrices become evident at higher $\epsilon = 0.6$. It is noteworthy that figure 3(e) shows the block structure for the US market and reveals the fact that behavior of US market was relatively calmer till 2002 and it became more volatile afterwards, the red–yellow stripes highlighting the crash periods. Similarly, figure 3(g) shows that the Japanese market became more volatile from 1990 onward; also, it went through more critical periods as compared to US market.

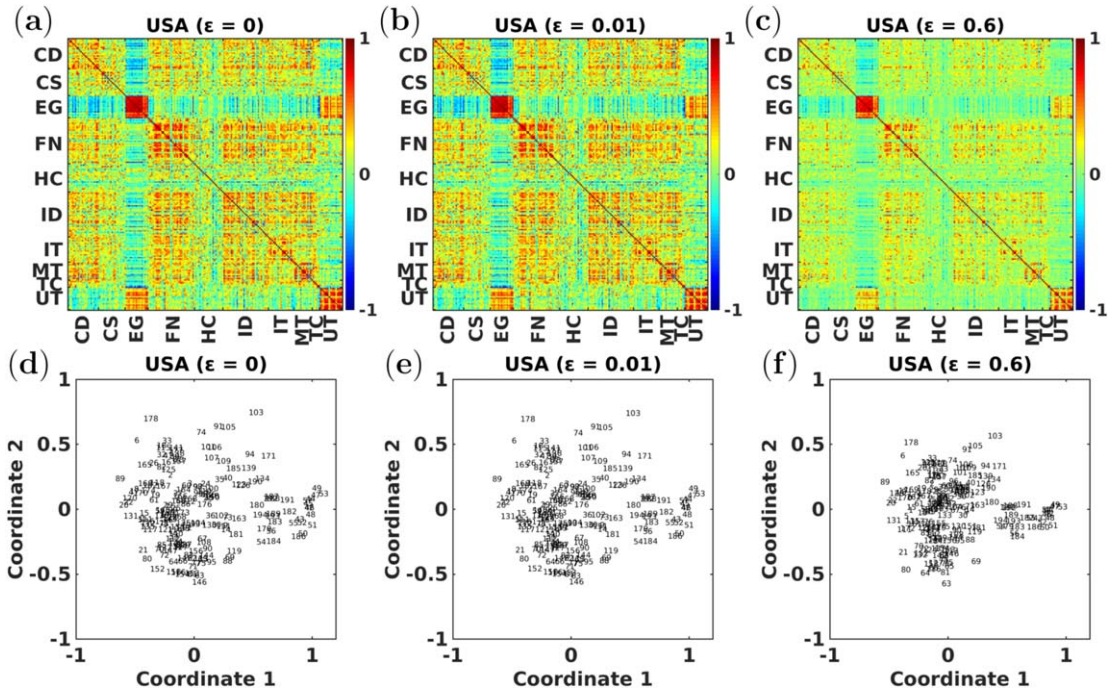

Figure 2. Noise-suppression in a short time cross-correlation matrix. (a) and (d) show the correlation pattern and corresponding MDS map, respectively, for a correlation matrix of short epoch of $M = 20$ d and $N = 194$ stocks of USA, ending on 30/11/2001. The power mapping method [19] is used to reduce the noise of the singular correlation matrix ($M < N$) formed by the short time series. Two different noise-suppression parameter values, $\epsilon = 0.01$ and $\epsilon = 0.6$, are used for this purpose. A small value of $\epsilon = 0.01$ is used for (b) and (e). In (c) and (f), when a higher distortion of $\epsilon = 0.6$ is given to the correlation matrix, the shape of emerging spectrum as well as the MDS map change drastically. In the MDS plot, the stocks with high correlations come nearer to each other and form more compact and distinct clusters, as compared to $\epsilon = 0$ and $\epsilon = 0.01$.

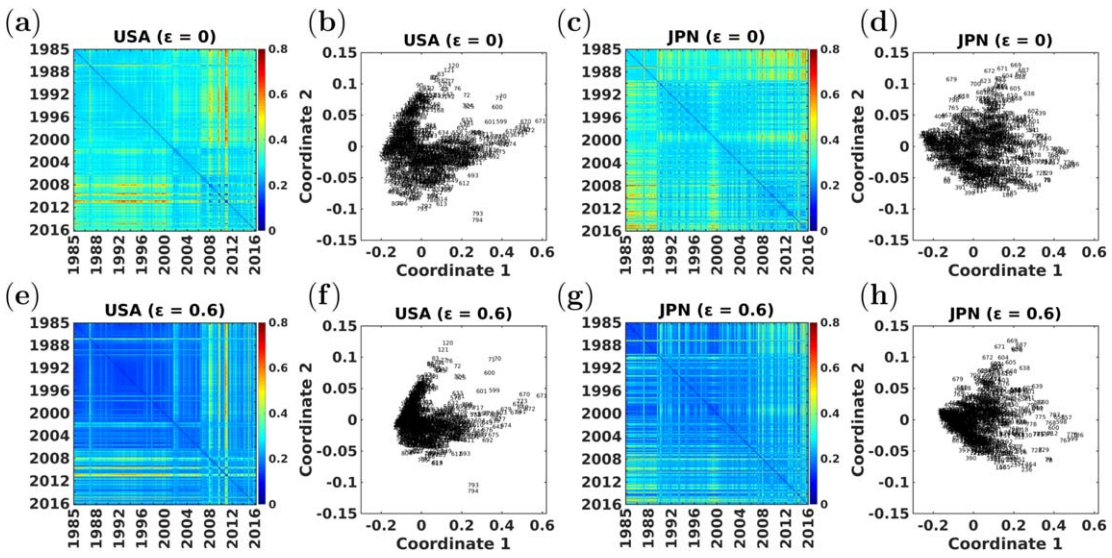

Figure 3. Noise-suppression in a similarity matrix (ζ) among correlation matrices over different epochs. (a) and (e) show the similarity matrices (without noise-suppression $\epsilon = 0$ and with noise-suppression $\epsilon = 0.6$, respectively) among 805 correlation matrices; (b) and (f) show the corresponding MDS maps for USA. (c) and (g) show the similarity matrices (without noise-suppression $\epsilon = 0$ and with noise-suppression $\epsilon = 0.6$, respectively) among 798 correlation matrices; (d) and (h) show the corresponding MDS maps for JPN. The similarity matrices give insight of the stock market evolution over 32 years (1985–2016). Red–yellow strips in the similarity matrices exhibit the crashes of the respective markets. The effect of noise-suppression is visible on both similarity matrices as well as MDS maps.

Importantly, the MDS maps with the noise-suppression parameter $\epsilon = 0.6$ are more compact and denser, which lead to better clustering and determination of optimal number of markets states (see also supplementary figures S2 and S3).

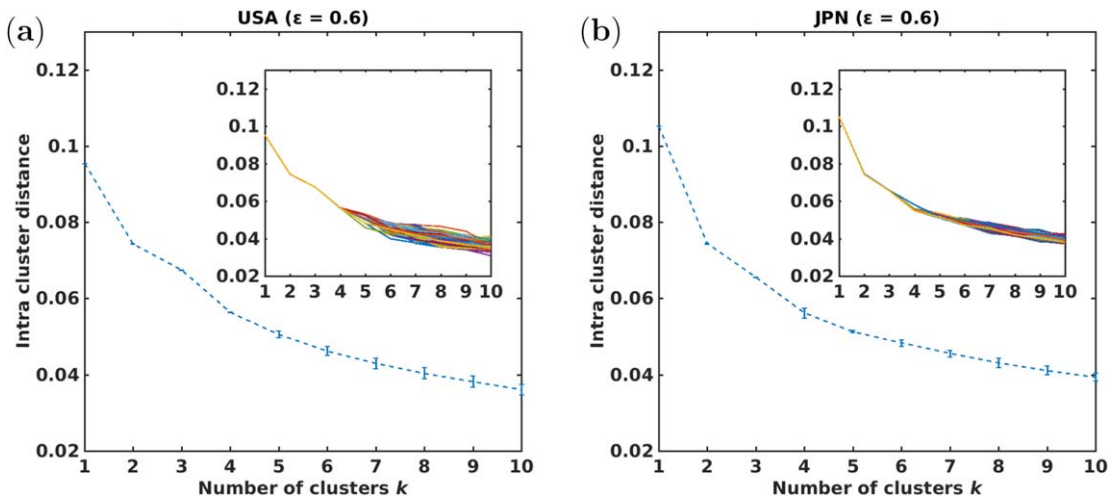

Figure 4. Plots of intra-cluster distance as a function of number of clusters. For (a) USA and (b) JPN, for the noise-suppression parameter $\epsilon = 0.6$. The k-means clustering is performed on the MDS map generated from 805 noise-suppressed correlation matrices of USA and 798 noise-suppressed correlation matrices of JPN, with 500 initial conditions in k-means clustering. The error bars are the standard deviations of the intra-cluster distances arising from the ensemble of 500 random initial conditions to centroids for the initial clustering. The plots show the minima of standard deviations at $k = 4$ for USA and $k = 5$ for JPN, which indicate the 'optimal' number of clusters. Inset: plot of intra-cluster distance versus k for all 500 random initial conditions. Each colored line corresponds to one such initial condition.

2.5. Determining optimal number of market states

To determine the number of market states, we find the number of clusters that can group together the noise-suppressed cross-correlation return matrices $\boldsymbol{C}(\tau)$ across different epochs $\tau = 1, \ldots, n$, based on their similarities [13]. We use the MDS map to visualize the information contained in $n \times n$ similarity matrix, and then use this MDS map with n objects for k-means clustering. The k-means clustering, which is a heuristic algorithm, aims to partition n numbers of correlation matrices into k clusters or groups in which each object/matrix belongs to the cluster with the centroid (nearest mean correlation), serving as a prototype of the cluster. In k-means clustering, the value of k can be optimized by different techniques [27, 28]. Here, we propose a new approach for optimizing k. We measure the mean and the standard deviation of the intra-cluster distances using an ensemble of fairly large number (say 500) of different initial conditions (choices of random coordinates for the k-centroids or equivalently random initial clustering of n objects); each set of initial conditions may result in slightly different clustering of the n different correlation matrices. If the clusters are distinct (or far apart in coordinate space) then even for different initial conditions, the k-means clusterings yield the same results, yielding a small variance of the intra-cluster distance. The problem of allocating the correlation matrices into the different clusters becomes acute when the clusters are close or overlapping, as the initial conditions can influence the final clustering. So there is a larger variance of the intra-cluster distance. Therefore, the minimum variance or standard deviation for a particular number of clusters displays the robustness of the clustering. For optimizing the number of clusters, we propose that one should look for *maximum* k, which has the *minimum variance* or standard deviation in the intra-cluster distances with different initial conditions. We propose this is easier than determining the 'elbow point' from the intra-cluster distance versus number of clusters curve [28].

For each cluster, one computes the average/variance of the point-to-centroid distances for all the points belonging to the cluster; the mean/variance of the intra-cluster distances is the mean/variance of the k values obtained from each of the k clusters. Next, we use 500 different initial conditions for the k-means clustering, each yielding a slightly different clustering result. One then computes the average as well as the variance (or standard deviation) of the mean intra-cluster distances among the ensemble of 500 runs. Then, the plots of average intra-cluster distance as functions of the number of clusters k for USA and JPN are shown in figure 4(a) and (b), respectively. The standard deviations of the intra-cluster distances measured for 500 initial conditions are shown as the error bars. The insets of figures 4(a) and (b), show the plots for 500 initial conditions. As mentioned earlier, the value of k is optimized by keeping the standard deviation lowest and the number of clusters highest; note that for $k = 1$, the standard deviations are always trivially zero. We find that for USA, the standard deviations are low till $k = 4$ and then grow for higher number of clusters; thus, $k = 4$ is the optimal number of clusters. For JPN, which is more complex than USA, the standard deviation is low for $k = 1, 2, 3$, increases for $k = 4$ and then decreases drastically for $k = 5$; beyond that again the standard deviation is higher. Thus, $k = 5$ is the optimum number of clusters for JPN.

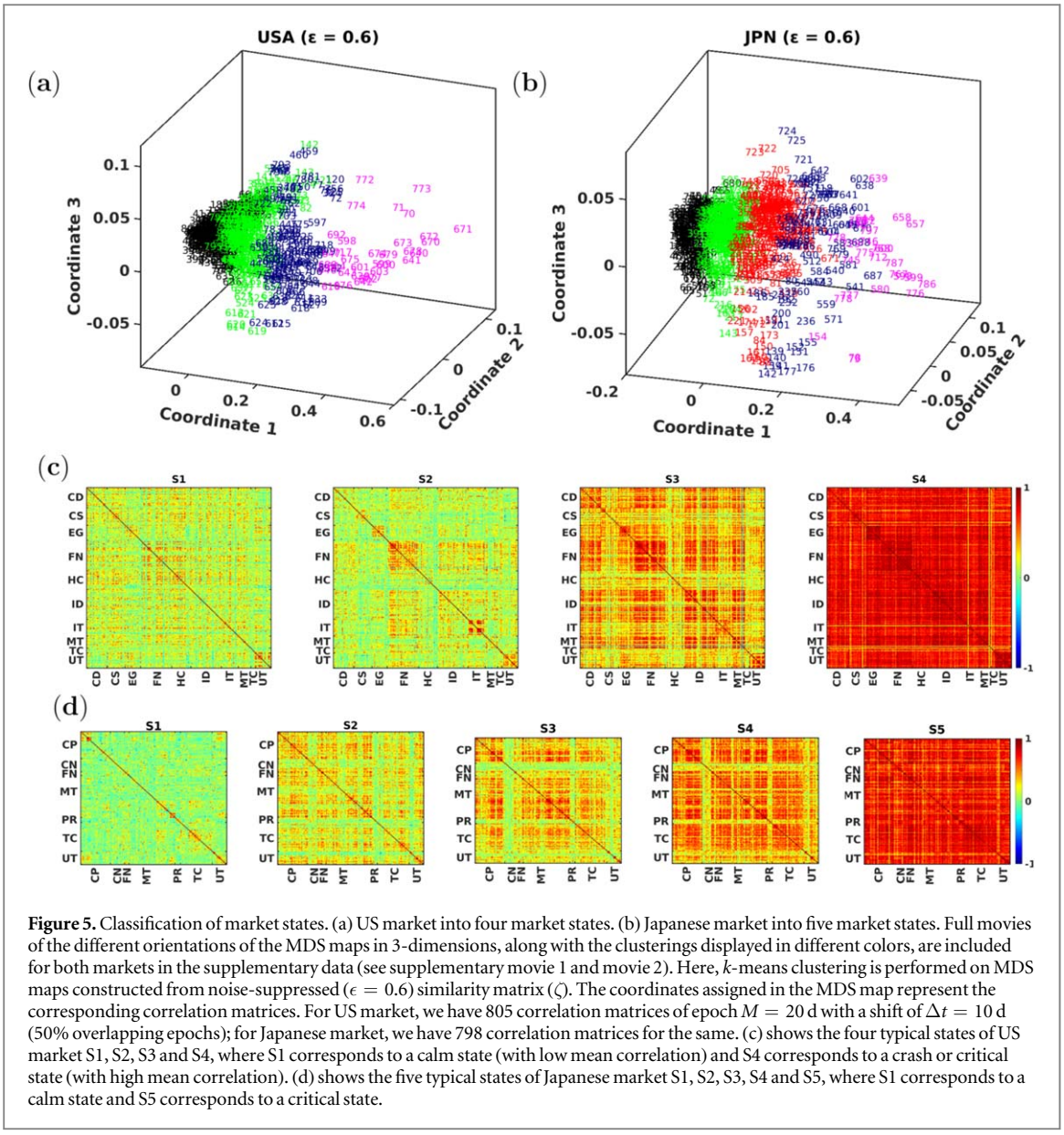

Figure 5. Classification of market states. (a) US market into four market states. (b) Japanese market into five market states. Full movies of the different orientations of the MDS maps in 3-dimensions, along with the clusterings displayed in different colors, are included for both markets in the supplementary data (see supplementary movie 1 and movie 2). Here, k-means clustering is performed on MDS maps constructed from noise-suppressed ($\epsilon = 0.6$) similarity matrix (ζ). The coordinates assigned in the MDS map represent the corresponding correlation matrices. For US market, we have 805 correlation matrices of epoch $M = 20$ d with a shift of $\Delta t = 10$ d (50% overlapping epochs); for Japanese market, we have 798 correlation matrices for the same. (c) shows the four typical states of US market S1, S2, S3 and S4, where S1 corresponds to a calm state (with low mean correlation) and S4 corresponds to a crash or critical state (with high mean correlation). (d) shows the five typical states of Japanese market S1, S2, S3, S4 and S5, where S1 corresponds to a calm state and S5 corresponds to a critical state.

The final k-means clustering of the correlation matrices in the similarity matrix is therefore performed for $k = 4$ clusters (USA) and $k = 5$ clusters (JPN), as shown in figures 5(a) and (b), respectively. (See supplementary movie 1 and movie 2 for more orientations of the MDS in 3-dimensions along with the clustering displayed in different colours, for USA and JPN, respectively.) We identify the points in each cluster (different colors represent different clusters) with similar correlation patterns and nearby mean correlation as one market state. Based on k-means clustering, figure 5(c) shows four different market states $S1$, $S2$, $S3$ and $S4$ of USA, where S1 corresponds to a calm state (with low mean correlation) and S4 corresponds to a crash or critical state (with high mean correlation); figure 5(d) shows five market states $S1$, $S2$, $S3$, $S4$ and $S5$ of JPN, where S1 corresponds to a calm state and S5 corresponds to a critical state, respectively. The states are arranged in the increasing order of mean correlation, in accordance with the finding in [13] that the clustering of correlation matrices yields similar (but not the same) results one would obtain by clustering just the highest eigenvalues. Here, we can also see clear differences structure-wise among the correlation matrices, e.g., there are strong intra-sectoral correlations within the energy, finance and utility sectors, in each of the market states of USA.

It may also be mentioned that the selection of noise-suppression parameter $\epsilon = 0.6$ is not totally arbitrary. We compared the plots of the average intra-cluster distance as function of the number of clusters for both USA and JPN, using ϵ ranging from 0.1 to 0.7 (shown in supplementary figures S2 and S3). The outcome of the comparison is that $\epsilon = 0.6$ yields the best results.

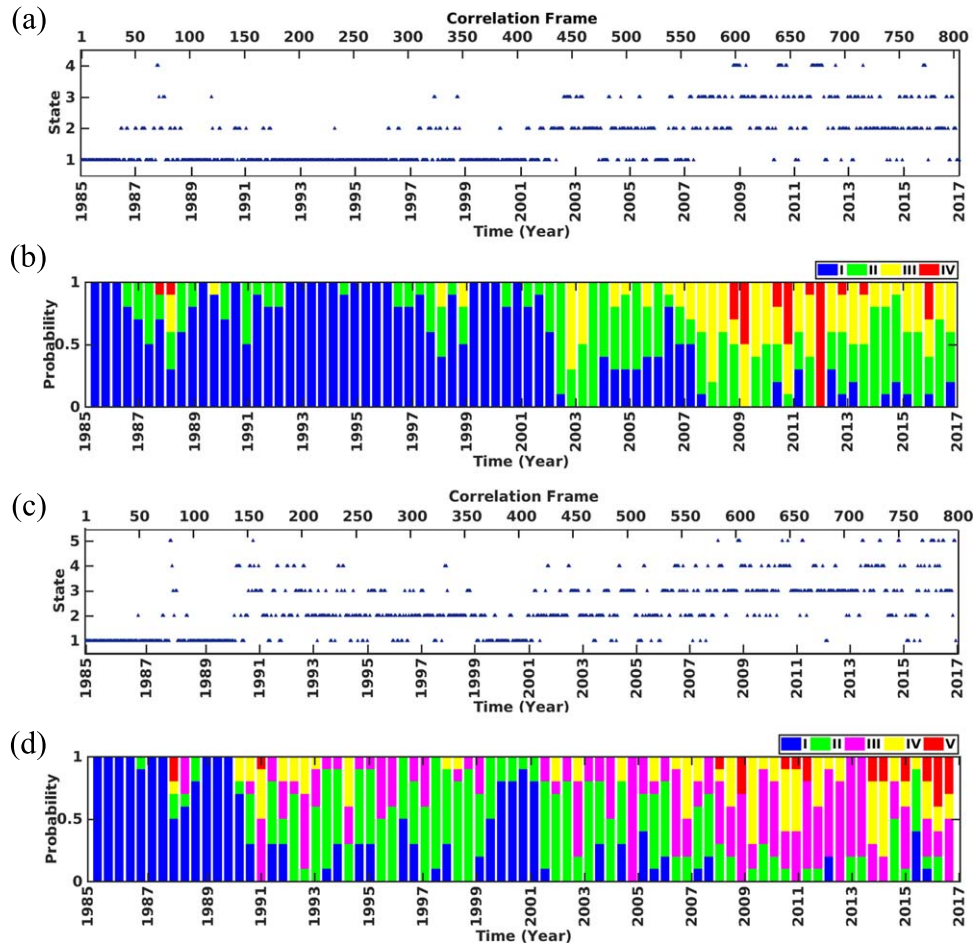

Figure 6. Dynamical evolution of market states for USA and JPN. (a) Temporal dynamics of the USA in four different states (*S*1, *S*2, *S*3 and *S*4) for the period of 1985–2016. (b) Probability plot of the four market states with each color length corresponds to the evolution probability of these four states during 110 d (10 overlapping epochs). (c) and (d) show similar results for JPN with five market states (*S*1, *S*2, *S*3, *S*4 and *S*5).

2.6. Transition probabilities and dynamics of market states

Once the classification of the short-time correlation matrices into different market states is complete, one can follow the evolution of the market as dynamical transitions between different markets states. Figures 6(a) and (c) show the evolution dynamics of market states of USA and JPN, during 1985–2016. In USA, the market oscillates among the four states *S*1, *S*2, *S*3 and *S*4. Often *S*1 or *S*2 states (with relatively low mean correlations) tend to remain in the same state for a long time; at other times, the market jumps to a higher mean correlation state *S*3 or *S*4. Similarly, for JPN the dynamical transitions among the five market states *S*1, *S*2, *S*3, *S*4 and *S*5. The probabilistic plots of the market states dynamics are shown in figures 6(b) and (d), for USA and JPN, respectively. The color length of any market state is the probability of that state computed during 110 d (10 overlapping epochs). Evident from the probability plots: (a) In USA, before 2002 the market was mostly in state *S*1; the market became more volatile, with more frequent transitions to other states, 2002 onward, and (b) in JPN, market became more volatile from 1990 onward. The same kind of behavior is also observed from the temporal evolution of the mean correlation (see supplementary figure S1).

Figures 7(a) and (b) show the bar plots of the counts for the co-occurrences of the market states for USA and JPN, respectively; the networks representing the transition probabilities for USA and JPN are, respectively, shown in figures 7(c) and (d), with corresponding values given in tables 2 and 3. For USA, the transition probability of $S3 \rightarrow S4$ is about 6%, $S3 \rightarrow S2$ is about 33%, and the probability of staying in the same state $S3 \rightarrow S3$ is about 58%. Thus, the state *S*3 warns of the possibility of a transition or acts like a 'precursor' to the state *S*4, though the probability for such a transition is still comparatively low. Similarly for JPN, for which the transition from $S4 \rightarrow S3$ (about 33%) is also quite a bit more probable than from $S4 \rightarrow S5$ (about 8%), the state *S*4 may act like a 'precursor' to the critical state *S*5. Entries just above and below the diagonals of the 3*D* bar plots are also quite high, which show that the transitions primarily happen between adjacent states. Exceptions of remote transitions occur, e.g., in the Black Monday crash of 1987.

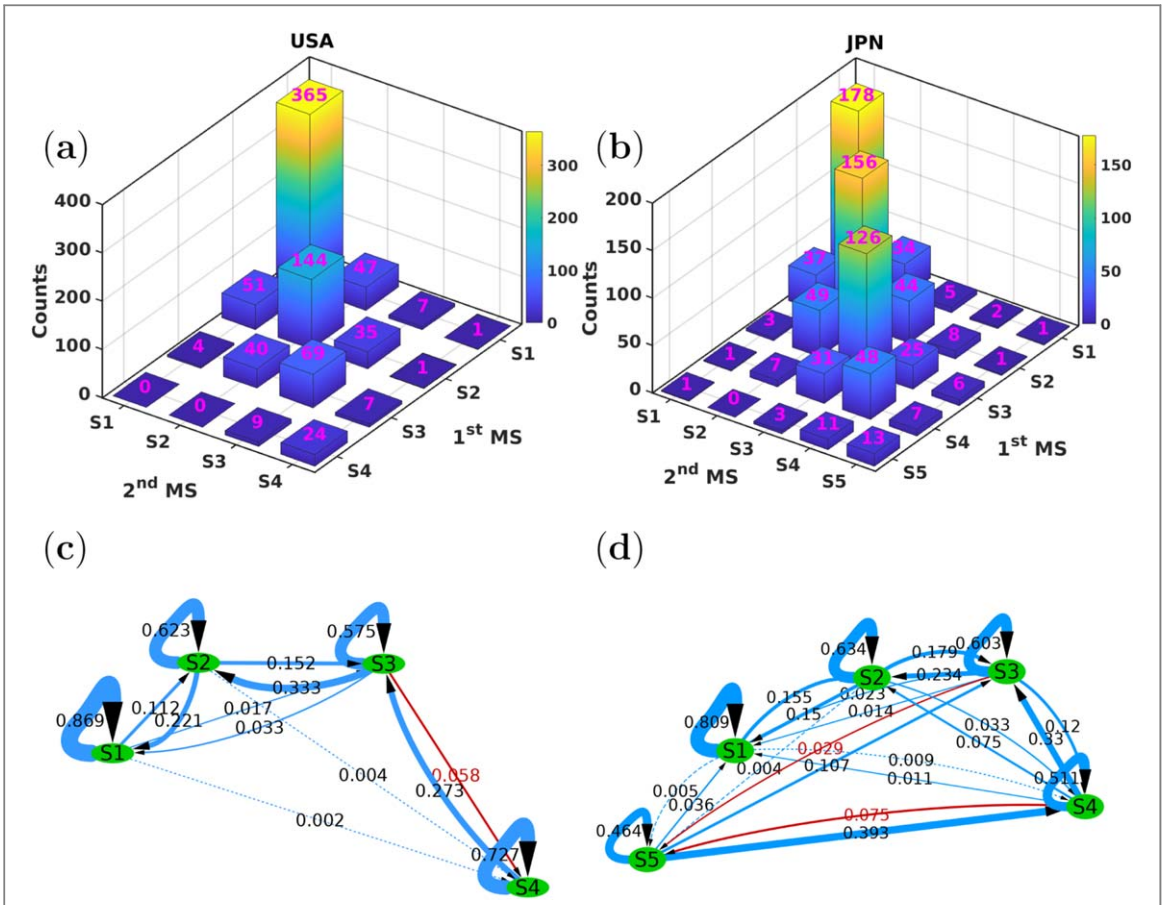

Figure 7. Transition probabilities of market states and determination of long-term 'precursors' of critical states. (a) and (b) 3*D* bar plots of counts of paired market states (MS) for USA and JPN, respectively; the sum of all the counts add up to 804 for USA and 797 for JPN, respectively, as they ought to be. (c) and (d) represent the networks of transition probabilities between different states for USA and JPN, respectively. See tables 2 and 3 for the transition probabilities. Note that the probability $S3 \rightarrow S4$ of USA is about 6%, which indicates that state *S*3 of USA may warn of a market crash *S*4; similarly, for JPN, the probability $S4 \rightarrow S5$ is about 8%, which indicates that the *S*4 state of JPN may warn of a critical state *S*5 (crash).

Table 2. USA: transition probabilities four market states (MS) (first is followed by second). Note that the numerical values given in the table are rounded off to 3 decimal places.

2nd MS → 1st MS ↓	S1	S2	S3	S4
S1	0.869	0.112	0.017	0.002
S2	0.221	0.623	0.152	0.004
S3	0.033	0.333	0.575	0.058
S4	0	0	0.273	0.727

Table 3. JPN: transition probabilities of five market states (MS) (first is followed by second). Note that the numerical values given in the table are rounded off to 3 decimal places.

2nd MS → 1st MS ↓	S1	S2	S3	S4	S5
S1	0.809	0.155	0.023	0.009	0.005
S2	0.150	0.634	0.179	0.033	0.004
S3	0.014	0.234	0.603	0.120	0.029
S4	0.011	0.075	0.330	0.511	0.075
S5	0.036	0	0.107	0.393	0.464

Finally, let us test the simple hypothesis whether the system jumps *randomly* from state S_i to S_j with probabilities W_{ij} or not. Note that, if we simply look at the curves in figures 7 (c) and (d), it is not obvious that this is indeed the case. However, if we make this hypothesis, we can obtain expressions for the probability that the system should be in one state over long times. This follows from the general theory of Markov chains [29], but for the sake of keeping the paper self-contained, we briefly explain the details below.

Let $P_i(n)$ be the probability that the system be in state i after n steps (epochs). Using the definition of W_{ij}, as well as the assumption that the transition to j depends only on the previous state via W_{ij}, and in no way on the previous history, we obtain

$$P_i(n+1) = \sum_j W_{ji} P_j(n), \tag{1}$$

where the sum is over all possible states j. After long times, it is plausible, and can in fact be proved rigorously, that the probability distribution becomes independent of n; in other words, the distribution reaches an equilibrium state $P_i^{(0)}$. The latter then satisfies the equations

$$P_i^{(0)} = \sum_j W_{ji} P_j^{(0)}. \tag{2}$$

This can be solved explicitly, if W_{ij} is known. The solution can be proved to be always positive, and can always be normalized such that

$$\sum_i P_i^{(0)} = 1, \tag{3}$$

so that the numbers $P_i^{(0)}$ can indeed be interpreted as a set of probabilities.

In the cases where the W_{ij}'s are given by table 2 (for USA) or table 3 (for JPN), it is straightforward to compute the equilibrium distributions: for USA, one finds:

$$P_1^{(0)} = 0.523 \qquad P_2^{(0)} = 0.288 \qquad P_3^{(0)} = 0.149 \qquad P_4^{(0)} = 0.040. \tag{4}$$

For JPN, on the other hand:

$$\begin{aligned} &P_1^{(0)} = 0.274 \qquad P_2^{(0)} = 0.308 \qquad P_3^{(0)} = 0.263 \\ &P_4^{(0)} = 0.119 \qquad P_5^{(0)} = 0.036. \end{aligned} \tag{5}$$

The actual frequencies for the four characteristic market states $S1$, $S2$, $S3$, and $S4$ of USA, obtained from figure 6(a), enable us to compute the probabilities: 0.523, 0.287, 0.149, and 0.041, respectively. Similarly, actual frequencies for the five characteristic market states $S1$, $S2$, $S3$, $S4$ and $S5$ of JPN, obtained from figure 6(c), enable us to compute the probabilities: 0.277, 0.308, 0.262, 0.118 and 0.035, respectively. These probabilities are indeed close to those in equations (4) and (5), and therefore our hypothesis is correct.

3. Summary and concluding remarks

In summary, we have studied the identification of market states and long-term precursors to critical states (crashes) in financial markets, based on the probabilistic occurrences of correlation patterns, determined using noise-suppressed short-time correlation matrices. We analyzed and compared the data of the S&P 500 (USA) and Nikkei 225 (JPN) stock markets over a 32 year period. We used the power mapping method to reduce the noise of the singular correlation matrices and obtained distinct and denser clusters in the two/three-dimensional MDS maps. The effects are prominent also on the similarity matrices and the corresponding MDS maps. The evolution of the market can be followed by the dynamic transitions between the market states. Using MDS maps, we applied k-means clustering to divide the clusters of similar correlation patterns of different epochs into k groups or market states. We showed that based on the cluster radii we could have a fairly robust determination of the optimal number of clusters. In each market, the value of optimal number of clusters was chosen by keeping the standard deviation of the intra-cluster distance 'minimum' and number of clusters 'highest'. Thus, based on the modified prescription of finding similar clusters of correlation patterns, we characterized US market by four market states and Japanese market by five. One must mention that this method yields the correlation matrices that correspond to the critical states (or crashes). We have verified that these indeed correspond to well-known financial market crashes (some minor, some major); also, specifically studied the properties of the emerging spectrum and characterization of the critical states (catastrophic instabilities) in [10, 15]. We also analyzed the co-occurrence probabilities of the paired market states. We observed that the probability of remaining in the same state is much higher than the transition to a different state. It implies that market states also feel an 'inertia'—stay in the same states for a long time. Also, probable transitions are the nearest neighbor transitions and from the co-occurrence table we showed that the probability reduces fast if one moved away from the diagonal. Hence, the transitions to other states mainly occurred in immediately adjacent

states with a few rare transitions to the remote states. The state adjacent to the critical state (crash) may behave like a warning or a long-term precursor for the critical state, and this prescription could be helpful in constructing an early warning system for financial market crashes.

Acknowledgments

AC and KS acknowledge the support by grant number BT/BI/03/004/2003(C) of the Government of India, Ministry of Science and Technology, Department of Biotechnology, Bioinformatics division, University of Potential Excellence-II grant (Project ID-47) of JNU, New Delhi, and the DST-PURSE grant given to JNU by the Department of Science and Technology, Government of India. KS acknowledges the University Grants Commission (Ministry of Human Research Development, Govt. of India) for her senior research fellowship. HKP and RC are grateful for postdoctoral fellowships provided by UNAM-DGAPA. FL acknowledges support from the project UNAM-DGAPA-PAPIIT IN103017 and CONACyT CB-254515. AC, KS and THS acknowledge the support grant by CONACyT through Project FRONTERAS 201, and also support from the project UNAM-DGAPA-PAPIIT IG 100616.

ORCID iDs

Anirban Chakraborti https://orcid.org/0000-0002-6235-0204
Francois Leyvraz https://orcid.org/0000-0001-9269-1248

References

[1] Vemuri V 1978 *Modeling of Complex Systems: An Introduction* (New York: Academic)
[2] Gell-Mann M 1995 *Complexity* **1** 16–9
[3] Bar-Yam Y 2002 *Encyclopedia of Life Support Systems (EOLSS)* (Oxford: UNESCO, EOLSS Publishers)
[4] Mantegna R N and Stanley H E 2007 *An Introduction to Econophysics: Correlations and Complexity in Finance* (Cambridge: Cambridge University Press)
[5] Bouchaud J P and Potters M 2003 *Theory of Financial Risk and Derivative Pricing: from Statistical Physics to Risk Management* (Cambridge: Cambridge University Press)
[6] Sinha S, Chatterjee A, Chakraborti A and Chakrabarti B K 2010 *Econophysics: An Introduction* (New York: Wiley)
[7] Chakraborti A, Muni Toke I, Patriarca M and Abergel F 2011 *Quant. Finance* **11** 991–1012
[8] Chakraborti A, Muni Toke I, Patriarca M and Abergel F 2011 *Quant. Finance* **11** 1013–41
[9] Chakraborti A, Challet D, Chatterjee A, Marsili M, Zhang Y C and Chakrabarti B K 2015 *Phys. Rep.* **552** 1–25
[10] Chakraborti A, Sharma K, Pharasi H K, Das S, Chatterjee R and Seligman T H 2018 arXiv:1801.07213
[11] Sornette D 2004 *Why Stock Markets Crash: Critical Events in Complex Financial Systems* (Princeton, NJ: Princeton University Press)
[12] Buchanan M 2000 *Ubiquity: Why Catastrophes Happen* (New York: Three Rivers Press)
[13] Münnix M C, Shimada T, Schäfer R, Leyvraz F, Seligman T H, Guhr T and Stanley H E 2012 *Sci. Rep.* **2** 644
[14] Chetalova D, Schäfer R and Guhr T 2015 *J. Stat. Mech.* **2015** P01029
[15] Pharasi H K, Sharma K, Chakraborti A and Seligman T H 2018 Complex market dynamics in the light of random matrix theory *New Perspectives and Challenges in Econophysics and Sociophysics* ed F Abergel *et al* (Milan: Springer)
[16] Vinayak, Schäfer R and Seligman T H 2013 *Phys. Rev.* E **88** 032115
[17] Laloux L, Cizeau P, Bouchaud J P and Potters M 1999 *Phys. Rev. Lett.* **83** 1467
[18] Plerou V, Gopikrishnan P, Rosenow B, Amaral L A N and Stanley H E 1999 *Phys. Rev. Lett.* **83** 1471
[19] Guhr T and Kälber B 2003 *J. Phys. A: Math. Gen.* **36** 3009
[20] Bouchaud J P and Potters M 2000 *Theory of Financial Risks* (Cambridge: Cambridge University Press)
[21] Schmitt T A, Schäfer R, Wied D and Guhr T 2016 *Empir. Econ.* **50** 1091–109
[22] Vinayak and Seligman T H 2014 *AIP Conf. Proc.* **1575** 196
[23] Borg I and Groenen P 1997 *Modern Multidimensional Scaling: Theory and Applications* (*Springer Series in Statistics*) (Berlin: Springer)
[24] 2017 Yahoo finance database.accessed on 7 July, 2017, using the r open source programming language and software environment for statistical computing and graphics URL https://finance.yahoo.co.jp/
[25] Mantegna R N 1999 *Eur. Phys. J.* B **11** 193–7
[26] Schäfer R, Nilsson N F and Guhr T 2010 *Quant. Finance* **10** 107–19
[27] Gonzalez T F 1985 *Theor. Comput. Sci.* **38** 293–306
[28] Bholowalia P and Kumar A 2014 *Int. J. Comput. Appl.* **105** 17–24
[29] Ross S M 1996 *Stochastic Processes* (New York: Wiley)