Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

Stability and hierarchy of quasi-stationary states: financial markets as an example

Yuriy Stepanov[1], Philip Rinn[2], Thomas Guhr[1], Joachim Peinke[2] and Rudi Schäfer[1]

[1] Faculty of Physics, University of Duisburg-Essen, Duisburg, Germany
[2] Institute of Physics and ForWind, Carl-von-Ossietzky University Oldenburg, Oldenburg, Germany
E-mail: yuriy.stepanov@uni-due.de

Received 22 March 2015
Accepted for publication 21 July 2015
Published 19 August 2015

Online at stacks.iop.org/JSTAT/2015/P08011
doi:10.1088/1742-5468/2015/08/P08011

Abstract. We combine geometric data analysis and stochastic modeling to describe the collective dynamics of complex systems. As an example we apply this approach to financial data and focus on the non-stationarity of the market correlation structure. We identify the dominating variable and extract its explicit stochastic model. This allows us to establish a connection between its time evolution and known historical events on the market. We discuss the dynamics, the stability and the hierarchy of the recently proposed quasi-stationary market states.

Keywords: correlation functions (experiments), critical phenomena of socio-economic systems, nonlinear dynamics, stochastic processes

J. Stat. Mech. (2015) P08011

© 2015 IOP Publishing Ltd and SISSA Medialab srl
1742-5468/15/P08011+19$33.00

Contents

1. Introduction

Big data is the buzzword of recent years, reflecting an ever increasing amount of electronically available data that demands analysis and interpretation. Our focus is on complex dynamical systems such as financial markets, where huge data sets exist in the form of multivariate time series. The dynamical behavior of such systems may reduce their complexity by self-organization [1]. System variables, which are measured as single time series, couple together to a few dominating variables, which accurately describe the system dynamics and allow for predictions. The self-organization may produce patterns in observed data which are generally difficult to uncover. A wide range of data analysis techniques is available and widely used, including graph theoretical information filtering [2–7], data clustering [8–13] and geometric approaches [14–18]. All these techniques are based on a similarity measure between the data points. There is a major disadvantage in this approach: The time information of the measured data is neglected. Thus, the system dynamics is not explicitely taken into account. On the other hand, dynamical variables of complex systems have been successfully described

doi:10.1088/1742-5468/2015/08/P08011

J. Stat. Mech. (2015) P08011

by stochastic processes [1, 19–22]. In this description the variables evolve in time according to deterministic dynamics, which gives access to system stability and fixed points and is exposed to generally non-trivial stochastic fluctuations. Here, we combine the data set analysis with stochastic methods in order to capture the full dynamics of the system. Similar techniques have proven successful in the description of complex dynamical systems [23–25]. We apply our approach to stock market data and investigate the dynamics of the stock price cross correlation, which is known to be non-stationary [26–29]. The paper is organized as follows: we present the data set and perform a geometric data analysis to uncover the dominating variable in section 2. In section 3 we identify the quasi-stationary states of the financial market following [12]. We draw connections to known historical events. In the present work we go beyond the pure identification of the quasi-stationary market states and investigate their stability and dynamics. We present the stochastic analysis in section 4 and discuss our results in section 5.

2. Analyzed data

In section 2.1 we introduce our data set and the analyzed quantities. We perform a geometric analysis of the data in section 2.2.

2.1. Observed quantities

We analyze daily adjusted closing stock prices $S_i(t)$ $i = 1, ..., K$ of the $K = 307$ companies in the S&P500 Index continuously traded over the period of 21 years ranging from early 1992 to the end of 2012. The data is freely available at finance.yahoo.com. To measure the correlations, we use the daily returns

$$r_i(t) = \frac{S_i(t+1) - S_i(t)}{S_i(t)} \tag{1}$$

and normalize them locally [30], to smooth out trends on very short times. We measure the time t in trading days. We then calculate the $K \times K$ correlation matrices $C(t)$ by averaging over a time window of $T = 42$ which is moved in one-day steps through the data. The elements of $C(t)$ are the Pearson correlation coefficients

$$C_{ij}(t) = \frac{\langle r_i r_j \rangle_T(t) - \langle r_i \rangle_T(t) \langle r_j \rangle_T(t)}{\sigma_i^{(T)}(t) \sigma_j^{(T)}(t)}. \tag{2}$$

Here $\sigma_i^{(T)}(t) = \sqrt{\langle r_i^2 \rangle_T(t) - \langle r_i \rangle_T^2(t)}$ is the time-dependent volatility and the sample average

$$\langle f \rangle_T(t) = \frac{1}{T} \sum_{s=t-T+1}^{t} f(s) \tag{3}$$

of a quantity $f(t)$ is evaluated over the T data points before t. We note that in contrast to the stock prices S_i and price returns r_i, the correlation coefficients $C_{ij}(t)$ are

J. Stat. Mech. (2015) P08011

bounded quantities. All together we obtain $N = 5169$ correlation matrices. The correlation matrices calculated on the short intervals T are noisy. We reduce the noise by averaging over the correlation coefficients which yields the mean correlation coefficient

$$\bar{c}(t) = \langle C(t) \rangle_{ij}. \tag{4}$$

Here $\langle ... \rangle_{ij}$ denotes the average over all $d = (K^2 - K)/2 = 46\,971$ independent correlation coefficients of every correlation matrix $C(t)$.

We recall the spectral decomposition

$$C(t) = \sum_{a=1}^{T} \lambda_a(t) \vec{u}_a(t) \vec{u}_a^{\dagger}(t) = \lambda_1(t) \left(\vec{u}_1(t) \vec{u}_1^{\dagger}(t) + \sum_{a=2}^{T} \mathcal{O}\left(\frac{\lambda_a(t)}{\lambda_1(t)} \right) \right) \tag{5}$$

of the $K \times K$ correlation matrix $C(t)$ [16, 17]. Here $\lambda_a(t)$ denotes the ath eigenvalue of $C(t)$, $\vec{u}_a(t)$ the corresponding normalized eigenvector and $\vec{u}_a^{\dagger}(t)$ its transpose. The rank of $C(t)$ is T and therefore only the first T eigenvalues are non-zero. For our data the first and the largest eigenvalue $\lambda_1(t) = \lambda_{\max}(t)$ is sufficiency larger than the other eigenvalues. All components of $\vec{u}_1(t)$ are approximately equal to 0.05, while the components of the other $T - 1$ eigenvectors spread around zero for every time t. Therefore $\vec{u}_1(t)$ corresponds to the dynamics of the whole market as in [16, 17]. Hence averaging over the correlation coefficients

$$\bar{c}(t) = \langle C(t) \rangle_{ij} \approx \kappa \lambda_{\max}(t) \tag{6}$$

we recover the largest eigenvalue. Here

$$\kappa = \langle \vec{u}_1(t) \vec{u}_1^{\dagger}(t) \rangle_{ij} \approx 229 \tag{7}$$

is an empirical factor which appears due to the noise in the data. The time evolution of the largest eigenvalue is strongly correlated with the mean correlation coefficient $\bar{c}(t)$, the Pearson correlation is 0.998. This connection has also been pointed out in [27]. Therefore, the quantities $\lambda_{\max}(t)$ and $\bar{c}(t)$ share the same dynamics. We will show in section 2.2 that $\bar{c}(t)$ has as much variability in the values as possible for our data. Figure 1(a) shows the time evolution of $\bar{c}(t)$. We also present the time evolution of the S&P500 Index in figure 1(b).

2.2. Geometric approach: principal component analysis

We identify each correlation matrix $C(t)$ with a correlation vector

$$\vec{c}(t) = \begin{bmatrix} c_1(t) \\ c_2(t) \\ \cdots \\ c_d(t) \end{bmatrix} \tag{8}$$

in the real d-dimensional Euclidian space $\mathbb{R}^d$. Here $c_i(t)$ is the ith component of $\vec{c}(t)$. We then apply the principal component analysis (Pearson [14], Hotelling [15]) to quantify orthogonal and therefore uncorrelated one-dimensional subspaces in our time series $c_i(t)$, $i = 1, ..., d$.

J. Stat. Mech. (2015) P08011

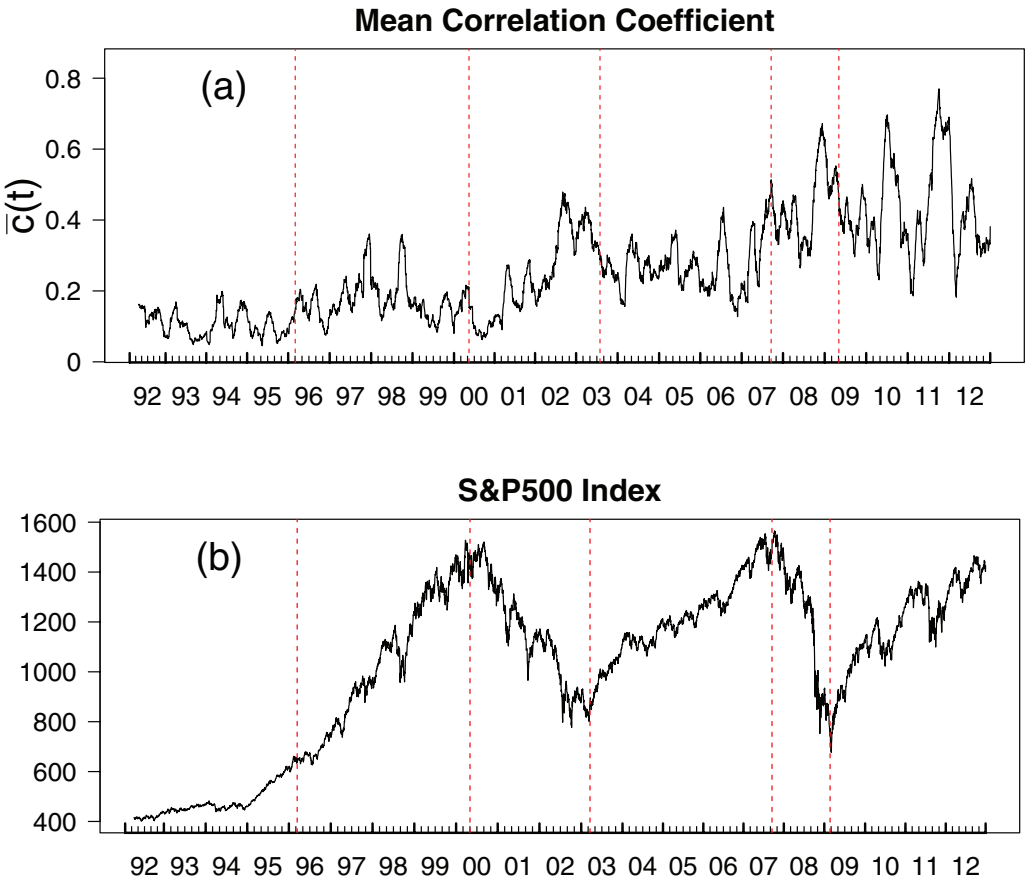

Figure 1. (a) Time evolution of the mean correlation coefficient $\bar{c}(t)$. (b) The S&P500 Index for the same time period. Dashed lines highlight economically distinct time intervals as described in section 3.2.

The first principal component is defined as the line in $\mathbb{R}^d$ with the largest possible variance of the data values. The other principal components are those with the largest data variance and orthogonal to the preceding components. The number of the principal components is smaller or equal to d. The principal components are spanned by the orthogonal eigenvectors $\hat{v}_i$, $i = 1, .., d$ of the symmetric $d \times d$ covariance matrix

$$W = AA^{\dagger}. \tag{9}$$

Here A is the $d \times T$ data matrix with d empirical times series $c_i(t)$ as rows and $A^{\dagger}$ denotes its transpose.

The rank of W is $\min(d, T)$ and we can not apply the PCA to our full data so we applied the principal component analysis (PCA) to randomly chosen 100 stocks ending up with $d = (100^2 - 100)/2 = 4950$ time series of length $T = 5169$. Figure 2(a) shows the eigen vector components distribution for the first ten principal components.

The components of the first normalized eigen vector are concentrated around a constant value 0.014, while the values of the other nine are symmetrically distributed around zero. Therefore the direction with the largest variance in data values is the subspace spanned by the vector

J. Stat. Mech. (2015) P08011

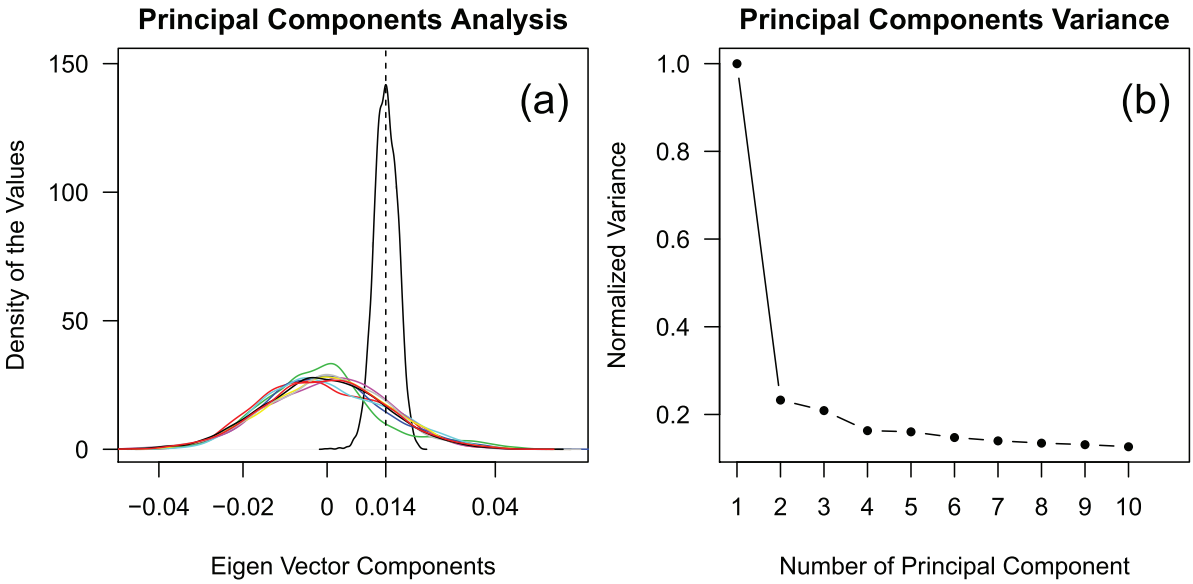

Figure 2. (a) The distribution of the first ten normalized principal components and (b) their variances normalized to the largest value.

$$\hat{v}_1 \equiv \frac{1}{\sqrt{d}}\begin{bmatrix}1\\1\\\cdots\\1\end{bmatrix} \approx \begin{bmatrix}0.014\\0.014\\\cdots\\0.014\end{bmatrix}. \tag{10}$$

The variance of data values for the first ten principal components are shown in figure 2(b). The variance of the first principal component is much larger than the others. The correlation matrices $C(t)$ from our data set seen as vectors $\vec{c}(t) \in \mathbb{R}^d$ are thus distributed along $\hat{v}_1$. Figure 3 shows the projection of our data onto the first three principal components in a scatter plot. The distribution of the data points along the first principal component is dominating. The contribution of the correlation matrix $C(t)$ to the first principal component at time t is given by the scalar product

$$\langle \vec{c}(t), \hat{v}_1 \rangle = \frac{1}{\sqrt{d}} \sum_{i=1}^{d} c_i(t) = \bar{c}(t)\sqrt{d}, \tag{11}$$

and turns out to be the mean correlation coefficient (4) times the fixed number $\sqrt{d}$. The dynamics of the market is therefore dominated by the movement along $\hat{v}_1$ which is given by $\bar{c}(t)$. Equation (11) confirms the spectral analysis results discussed in section 2.1. We note that spectral analysis of the correlation matrix $C(t)$ is the principal component analysis of the standardized returns

$$\hat{r}_i(t) = \frac{r_i(t) - r_i\rangle_T(t)}{\sigma_i^{(T)}(t)}. \tag{12}$$

treated as element of $\mathbb{R}^K$. Therefore the projection of (12) on the first principal component in $\mathbb{R}^K$ at time t is equal to the non-weighted average of $\hat{r}_i(t)$.

The projections $\langle \vec{c}(t), \hat{v}_2 \rangle$ and $\langle \vec{c}(t), \hat{v}_3 \rangle$ describe system dynamics along the second and third principal component and are shown in figure 4.

J. Stat. Mech. (2015) P08011

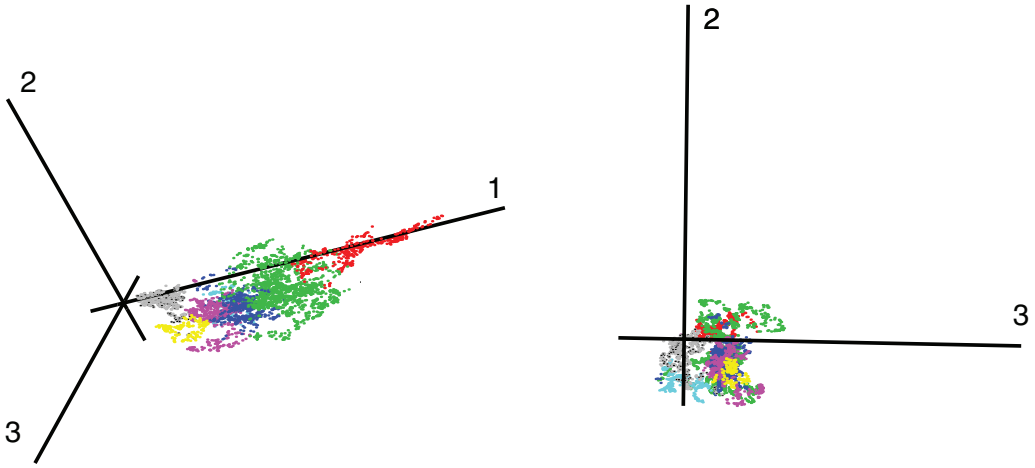

Figure 3. Data set projected onto the first three pricipal components. Different colors highlight different market states as explained in section 5.

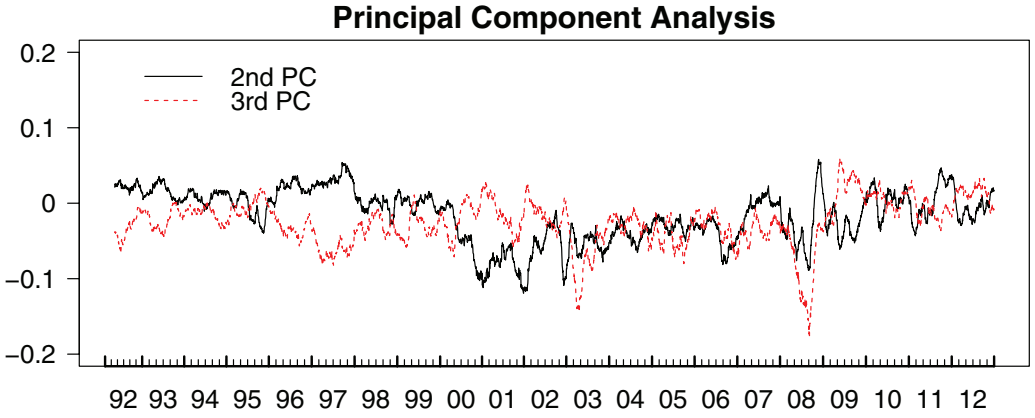

Figure 4. Time evolution of the projections $\langle \vec{c}(t), \hat{v}_2 \rangle$ (solid, black) and $\langle \vec{c}(t), \hat{v}_3 \rangle$ (dashed, red) onto the second and third principal components normalized by $\sqrt{d}$.

3. Market states: distinct periods of the market

We cluster the data following [12] and identify the quasi-stationary states of the financial market which we present in section 3.1. We connect the characteristic states on the market to the known historical events in section 3.2.

3.1. Market states

In the previous section we showed that our data is spread along a few dominating subspaces in $\mathbb{R}^d$. To quantify the similarity between any two correlation matrices $C(t_a)$ and $C(t_b)$ we calculate the distance

$$D_{ab} = \| C(t_a) - C(t_b) \| = \| \vec{c}(t_a) - \vec{c}(t_b) \| \tag{13}$$

via the Euclidean norm on $\mathbb{R}^d$ normalized by $\sqrt{d}$.

J. Stat. Mech. (2015) P08011

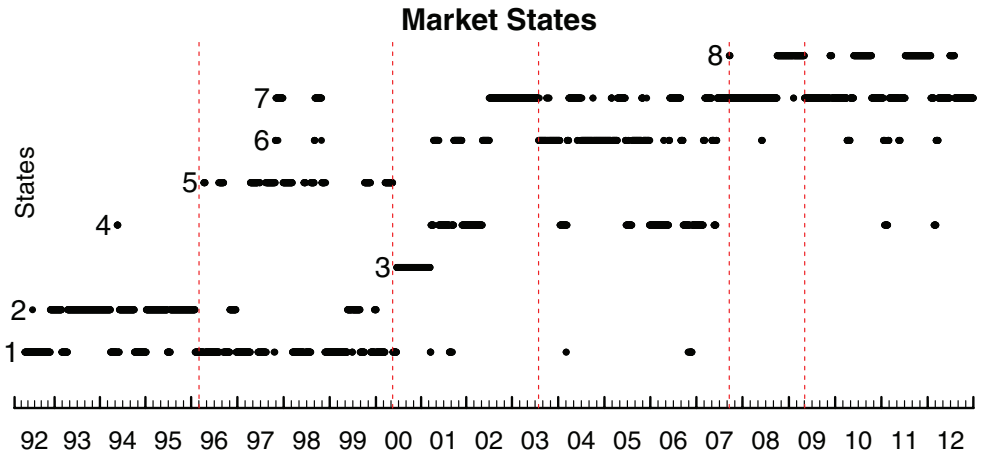

Figure 5. Time evolution of the market states. Dashed lines highlight economically distinct time intervals as described in section 3.2

As the next step we use the bisecting k-means clustering algorithm [9]. At the beginning of the clustering procedure all of the correlation matrices are considered as one cluster, which is then divided into two sub clusters using the k-means algorithm with $k = 2$. For each cluster α we then calculate its cluster center

$$\vec{\mu}_\alpha = \frac{1}{N_\alpha} \sum_{t \in \alpha} \vec{c}(t), \tag{14}$$

which is the mean correlation matrix in this cluster. Here N_α denotes the number of the cluster elements and $t \in \alpha$ symbolically denotes all times t for which $\vec{c}(t)$ is in the cluster α. The separation procedure is repeated until the cluster size

$$R_\alpha = \frac{1}{N_\alpha} \sum_{t \in \alpha} \|\vec{c}(t) - \vec{\mu}_\alpha\| \tag{15}$$

is smaller than a given threshold for every cluster α. We choose the mean distance to be smaller than 0.164 to achieve 8 clusters as in [12]. The market is said to be in a market state α at time t, if the corresponding correlation matrix $C(t)$, and hence the correlation vector $\vec{c}(t)$, is in the cluster α. The time evolution of the market states is shown in figure 5. In figure 6 the corresponding clustering tree is shown. The state occupied on the first day of our data is labeled by one. The remaining states are labeled according to the mean value of $\bar{c}(t)$ within the states as shown in figure 6. We group the states into three main classes. The market states one, two and three represent calm states. The states four, five and six are intermediate states. The states seven and eight are the turbulent states. The financial market evolves between these different states. New states form and existing states vanish in the course of time. For example the first four years are dominated by the states 1 and 2 in the last four years mainly the states 8 and 7 are occupied.

3.2. Distinct time periods

We divide the entire time period into six dynamically and economically distinct intervals.

J. Stat. Mech. (2015) P08011

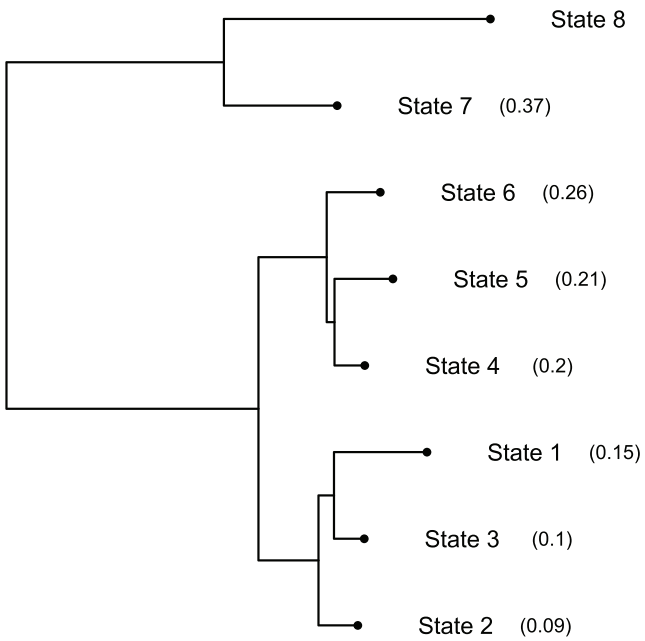

Figure 6. Clustering tree of the market states clustering. The mean value of $\bar{c}(t)$ within the states is given in parentheses.

(i) Early 1992 to spring of 1996: in this rather calm period $\bar{c}(t)$ varies between 0 and 0.2. The S&P500 Index continuously grows with moderate volatility. The market mainly occupies the first and the second state.

(ii) From spring 1996 until spring 2000: the range that $\bar{c}(t)$ explores as well as the S&P500 Index drastically increase. The volatility also becomes larger. The increase of $\bar{c}(t)$ is explained by the appearance of strongly correlated industrial sectors during this period, especially the technology sector. The market state two almost disappears and the market jumps mainly between states five and one. We note that the fifth state appears only during this period.

(iii) Spring 2000 to the second half of 2003: this period fully covers the *dot-com bubble* and is known as a very turbulent time in financial markets due to the crisis. The S&P500 Index drops continuously, losing about half of its value. The mean correlation coefficient reaches its maximum at 0.48. At the beginning of the crises state 3 appears for about one year. This state appeared only once during the entire time period. In the second half the market is switching between states four and six and occupies state seven by the end of 2002. This period includes the market response to the 9/11 attacks.

(iv) From the second half of 2003 until fall of 2007: this period covers the four years period before the recent global financial crises up to the 1 year period before the collapse of *Lehman Brothers*. As seen from the S&P500 Index in figure 1, the market seems to recover after the dot-com crisis but $\bar{c}(t)$ does not calm down and strongly fluctuates around a mean value 0.28. The market is jumping between states four, six and seven. State six is occupied mainly during this interval.

J. Stat. Mech. (2015) P08011

(v) From October 2007 until March 2009: this period covers the late-2000s financial crisis. The S&P500 Index drops continuously and looses approximately half of its value. The mean correlation coefficient is peaked sharply at 0.67. The market is mainly in state seven and occupies the eighth state by the end of 2008.

(vi) March 2009 to end 2012: the market seems to slowly recover as the S&P500 Index grows again. The growth interrupted by drastic drops. This is reflected in high peaks of $\bar{c}(t)$, which accounts its maximum value 0.77 in the analyzed 21 years. The mean correlation coefficient does not relax to the values it had before the crisis. The market is switching between states seven and eight and decays for short time into the states four and six.

4. Stochastic analysis

We describe the stochastic process used to model $\bar{c}(t)$ in section 4.1. In section 4.2 we explain how the explicit model is extracted form the time series. We describe the stochastic analysis of the market states in section 4.3.

4.1. Stochastic processes

We model $\bar{c}(t)$ as a stochastic process described by a Langevin equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\bar{c}(t) = f(\bar{c}, t) + g(\bar{c}, t)\Gamma(t), \tag{16}$$

i.e. a stochastic differential equation (SDE) for the variable $\bar{c}(t) \in \mathbb{R}$. Here f is the deterministic part of (16)—the drift function and g is the diffusion function, which defines the stochastic part of (16). $\Gamma(t)$ is the δ-correlated Gaussian white noise with $\langle\Gamma(t)\rangle = 0$ and $\langle\Gamma(t_1)\Gamma(t_2)\rangle = \delta(t_1 - t_2)$. We note that for the dimensionless variable $\bar{c}(t)$ the drift function has a dimension of inverse time and the diffusion function has a dimension of inverse square root of time.

The solution of (16) is defined in terms of stochastic integrals, which depend on the choice of the discretization [31–33]. Throughout this paper we use Itô's choice (see Itô's interpretation of SDEs [32, 34]). The advantage of Itô's definition is that the diffusion term g is uncorrelated with the Gaussian white noise $\langle g(\bar{c}, t)\Gamma(t)\rangle = 0$ [32]. The drift and diffusion terms can therefore be obtained as conditional moments [32, 35]

$$f(c, t) = \lim_{\tau\to 0} \frac{\langle \bar{c}(t+\tau) - \bar{c}(t)\rangle}{\tau}\bigg|_{\bar{c}(t)=c}, \tag{17}$$

$$g^2(c, t) = \lim_{\tau\to 0} \frac{\langle (\bar{c}(t+\tau) - \bar{c}(t))^2\rangle}{\tau}\bigg|_{\bar{c}(t)=c}. \tag{18}$$

Here c denotes the value of the stochastic variable $\bar{c}(t)$ at which the value of the drift or the diffusion is evaluated. At this one instant we distinguish between $\bar{c}(t)$ and a

J. Stat. Mech. (2015) P08011

particular numerical value c. The average in equations (17) and (18) is performed over all realizations of $\bar{c}(t)$ for which the condition $\bar{c}(t) = c$ holds. These equations express therefore the time derivative of the mean displacement and its square of $\bar{c}(t)$ at c.

Expressions (17) and (18) allow one to estimate the drift and diffusion directly from the empirical data as shown in [19, 20] and sketched below, see [1, 21, 22, 36] for applications. In the present work we model $\bar{c}(t)$ by an Itô stochastic process and estimate the deterministic as well as the stochastic part of the corresponding SDE from the empirical time series.

4.2. Estimation of the conditional moments

For the estimation of the drift and the diffusion directly from the data set we mainly follow [19–22]. Here, we briefly sketch the estimation procedure for the drift function, i.e. the first conditional moment (17), as the estimation of the diffusion function (18) works accordingly. We first introduce a new function

$$M_c(\tau) = \left.\frac{\langle \bar{c}(t+\tau) - \bar{c}(t)\rangle}{\tau}\right|_{\bar{c}(t)=c}, \tag{19}$$

for which the drift function

$$f(c,t) = \lim_{\tau\to 0} M_c(\tau) \tag{20}$$

is obtained at $\tau = 0$. We note that we dropped the time variable t in the argument of M in equation (19) for brevity. For the estimation of $M_c(\tau)$ at fixed c as a function of τ we divide the time series $\bar{c}(t)$ into bins with equal number of data points. For every bin I the function $M_c(\tau)$ is then estimated as

$$M_{\bar{c}_I}(\tau) = \left.\frac{\langle \bar{c}(t+\tau) - \bar{c}(t)\rangle}{\tau}\right|_{\bar{c}(t)\in I}. \tag{21}$$

Here $\bar{c}_I$ is the mean value of $\bar{c}(t)$ in bin I and the average is performed over all data in this bin. We note that for the empirical data this estimation can only be done for discrete values of $\tau = 1, 2, 3....$ We then fit a second order polynomial in τ to the empirically estimated values of (21), extracting the desired value of the drift at $\bar{c}_I$ as the constant coefficient of the fitted function. The estimation of (19) is only possible for the realized values of the empirical times series $\bar{c}(t)$.

Instead of analyzing the drift function (17) itself, it is more convenient to consider the potential function

$$V(\bar{c},t) = -\int^{\bar{c}} f(x,t)\mathrm{d}x, \tag{22}$$

defined as the negative primitive integral of f. The minus sign is a convention. The dynamics of the system is encoded in the shape of $V(\bar{c},t)$: the local minima of the potential function correspond to the quasi-stable equilibria, or quasi-stable fixed points, around which the system oscillates. In contrast, local maxima correspond to unstable fixed points. We note that potential functions are defined up to an additive constant.

J. Stat. Mech. (2015) P08011

For the dimensionless variable $\bar{c}(t)$ the dimension of the potential function is the inverse time.

4.3. Market states dynamics

To quantify the market dynamics while it is in a fixed market state α we restrict the estimation of (21) and evaluate only the data points

$$\{\bar{c}(t), \bar{c}(t+\tau) \mid t \in \alpha\} \tag{23}$$

for each state α. We therefore consider only displacements along the first principal component within the market states. No state transitions are allowed. Potential functions estimated this way provide information about the stability of the market states and reveal the fixed points.

As we mentioned in section 3.1 we group the states into the three main classes according to the hierarchical structure as shown in figure 6. We estimate the potential functions for each class A evaluating only the data points

$$\{\bar{c}(t), \bar{c}(t+\tau) \mid t \in A\}. \tag{24}$$

Here $t \in A$ symbolically denotes all time points at which market is in a state of the class A. For example the market might be in the state 1 at time t and in the state 2 at time $t+\tau$, as these two clusters belong to the same class. We therefore consider only displacements within A and allow for state transitions between state of the same class.

5. Results

We show the estimated diffusion function (18) in section 5.1 and discuss the estimated potential function (22) in section 5.2. In section 5.3 we take a closer look at the dot-com bubble. A detailed study of the market states dynamics is presented in section 5.4.

5.1. Diffusion term

To quantify the time dependency of the diffusion function $g(\bar{c}, t)$ we estimated the second conditional moment (18) on a time window of four trading years (1008 trading days) which is moved in steps of two trading months (42 trading days). All together we obtain 100 estimates for $g(\bar{c}, t)$ which we present in figure 7. As we explained in section 4.2, the estimation is only possible for the realized values of $\bar{c}(t)$. We therefore put all estimated values in a single diagram. We then fit the estimated values by the time-independent function

$$g(\bar{c}) = \lambda\sqrt{(\bar{c} - c_{\min})(c_{\max} - \bar{c})}\,, \tag{25}$$

which fits our data well, see figure 7. The diffusion function (25) is widely used to model the stochastic correlation [37–41], as it limits the values of the correlation to the range $[c_{\min}, c_{\max}]$. From the estimated parameters

J. Stat. Mech. (2015) P08011

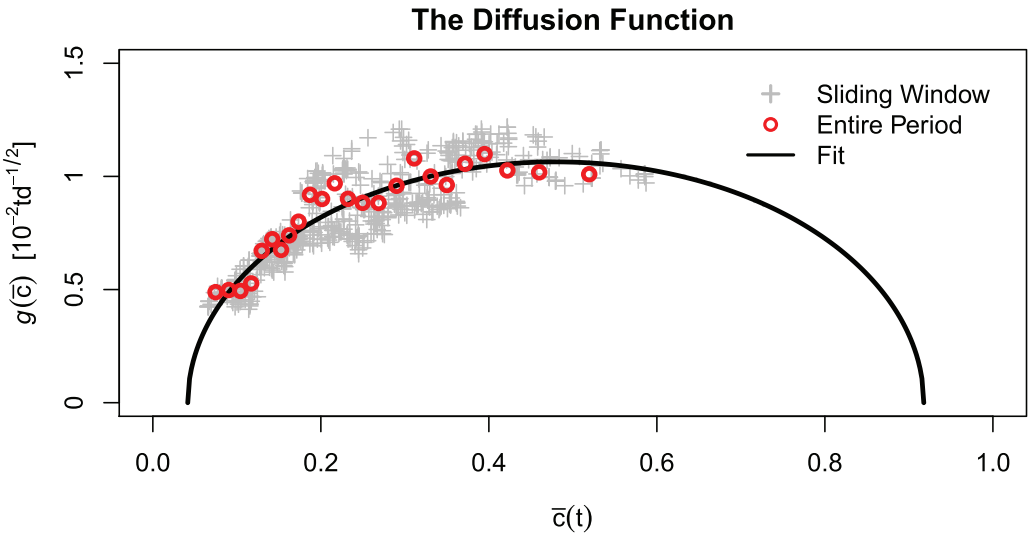

Figure 7. The diffusion function estimated on the sliding window (crosses (+)). The circles (○) show the estimated diffusion function on the entire time period at once. The solid curve shows the fitted function (25). We only use values estimated on the sliding windows for the fit.

$$\lambda = 0.0245\,\mathrm{td}^{-1/2}, \tag{26}$$

$$c_{\min} = 0.042, \tag{27}$$

$$c_{\max} = 0.918, \tag{28}$$

we obtain the characteristic time scale of the system

$$t_0 = \frac{1}{\lambda^2} = 1666 \text{ trading days}, \tag{29}$$

which turns out to be approximately 5.6 trading years. References [27, 28] show that the correlation between industry indices has both a fast and a slow dynamics. While the time scale of the fast dynamics is the monthly scale, the slow dynamics time scale is at least as slow as five years. This time scale perfectly agrees with the value we find for the characteristic time scale (29). For consistency we estimate (18) for the entire time series $\bar{c}(t)$ at once, as shown in figure 7. We note that we fitted (25) only to the data obtained on the sliding window.

5.2. Time evolution of the potential functions in the entire time period

To quantify the time dependence of the drift function $f(\bar{c}, t)$ we estimate the first conditional moment (17) on a time window of four trading years (1008 trading days) which is moved in steps of two trading months (42 trading days). All together we obtain 100 estimates for $f(\bar{c}, t)$. We then calculate the potential functions (22) which are presented

J. Stat. Mech. (2015) P08011

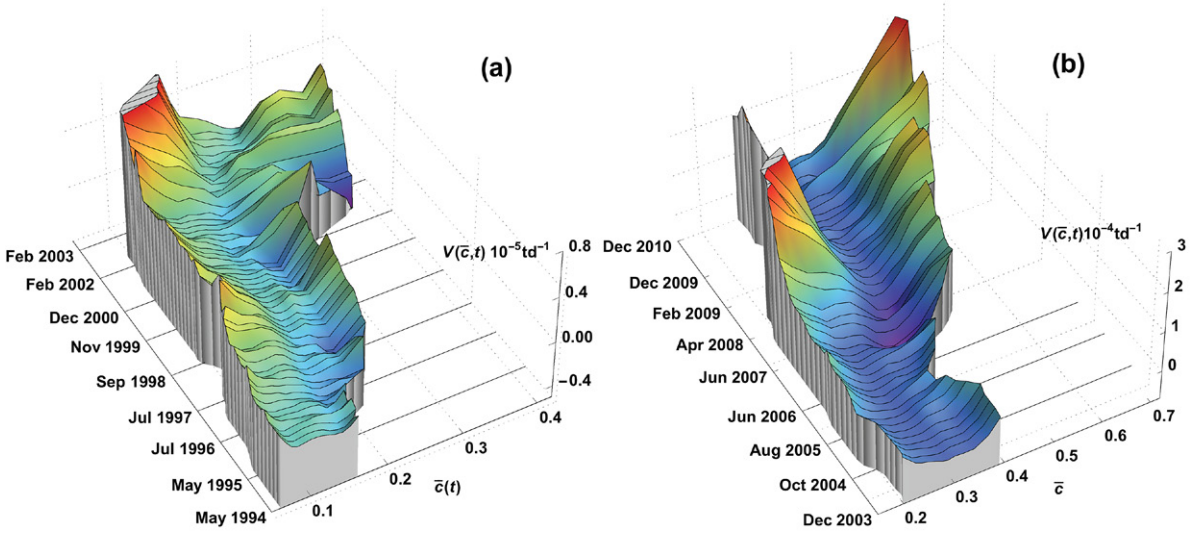

Figure 8. Time evolution of the potential function (22) from early 1992 to the end of 2004 (a) and from 2002 to the end of 2012 (b) estimated on a time window of four trading years which is moved in steps of two trading months. The dates mark the time points in the middle of the estimation time windows. Representation is according to equation (30).

in figures 8(a) and (b). The dates mark the time points in the middle of the estimation time windows. In contrast to the diffusion function, the drift function turns out to be time-dependent. Therefore it is difficult to graphically present many curves in a single diagram, as the potential function (22) is defined up to an additive constant. To work around this problem we set

$$V(\bar{c}_0, t) = 0, \tag{30}$$

where $\bar{c}_0$ denotes the value at which $V(\bar{c}, t)$ has its minimum in the first half of its values. In this representation the deeper a potential function is, the higher are the boundaries.

Figure 8(a) shows the results from early 1992 to the end of 2004. The distinct time periods described in section 3.2 are clearly recognizable in the shape of the potential function. It is flat and approximately constant at the beginning. It gets deeper in the middle of the period during the turbulent time in 1997–98. The two local minima show instabilities on the market. By the end of the period the dot-com crises is reflected in the shape of $V(\bar{c}, t)$. Its boundaries get higher and it has many deep minima at high values of $\bar{c}(t)$.

Figure 8(b) shows the results from early 2002 to the end of 2012. Similar to the previous case, $V(\bar{c}, t)$ is flat and constant during the relatively calm period in the early 2000s. It changes its shape drastically in the second half of the 2007 and gets a deep local minimum around $\bar{c}(t) \approx 0.4$. The boundaries get very high during the late-2000s financial crisis. We note that $V(\bar{c}, t)$ does not become flat after 2010.

We showed that $\bar{c}(t)$ is described by a stochastic process (16) with a time-independent diffusion term and a time-dependent drift function. In section 2 we showed that the mean correlation coefficient is the dominating variable of the collective market dynamics. The non-stationarity of the potential function is therefore explained by deterministic changes in the collective correlation structure on the market.

J. Stat. Mech. (2015) P08011

5.3. Zooming into the dot-com bubble

In the previous section we showed that the market evolves in time, switching back and forth between different market states. As an example of a state transition we estimate $V(\bar{c}, t)$ in the period from early 1999 to early 2006. The interval covers the dot-com bubble. To achieve higher time resolution we perform the estimation on a time window of two trading years (512 trading days), sliding it in steps of one trading month (21 trading days). Figure 9 shows the time evolution of the estimated potential function. It is flat at the beginning where the market is mainly in the states 1 and 2, see figure 5. During the crisis the values of $\bar{c}(t)$ increase. Therefore the estimated potential function moves along the $\bar{c}$ axis. A deep minimum builds up and the boundaries get higher. The market jumps through the states 3, 4 and 6, ending up in state 7 by the end of 2002. By the end of 2003 the market settles into state 6 with only short jumps into the states 4 and 1. The potential function becomes constant but has changed its shape compared to the pre-crisis period. The market therefore jumps from a stable state to a turbulent state and then down to another stable state.

5.4. Market states dynamics: stability, hierarchy and state transition

In the previous sections we showed that the mean correlation coefficient is described by a stochastic process (16) with the time-independent diffusion function (25) and the time-dependent drift function. Especially, calm and turbulent periods can be distinguished by the shape of $V(\bar{c}, t)$. To quantify the market dynamics in a given market state we estimate the potential function for the data points (23). Thus, we only accounted for displacements within a fixed market state.

The time series for the states 3, 4 and 5 are too short for the estimation of (17), so we combined the time series of the states 2 and 3 together as well as 4 and 5. We denote the resulting states by 2 + 3 and 4 + 5 respectively. As shown in figure 6, these pairs consist of states of the same class. Figures 10(a)–(c) shows the resulting potential functions for each market state.

Potential functions provide information about the stability of market states. This notion of the stability is not due to the time which the market spends occupying a certain state, but is given by the dynamics of the market. States 1, 2 + 3, 6 and 8 are stable states, as their potential functions have a single deep minimum and therefore a clearly defined fixed point. State 8 mainly appears during the latest financial crisis and represents a strong collective correlation on the market. In contrast state 7 is very unstable. Not only has its potential function two local minima, but it is also the deepest one. The correlation structure is non-stationary within the market state 7. The combined state 4 + 5 has a half-open potential function. States 4 and 5 are intermediate states between calm and turbulent periods, see figure 5. We note that within stable states $\bar{c}(t)$ is described by SDE (16) with the diffusion function (25) and a linear drift function.

In section 3.1 we grouped the market states into three classes according to the clustering tree, see figure 6. Not all of the market states appear simultaneously in a given time interval, as shown in figure 5. The first four years of the analyzed time period are dominated by the states 1 and 2, which belong to the first class. In the last four years basically only the states 7 and 8 appear, which build the third class. To quantify the

J. Stat. Mech. (2015) P08011

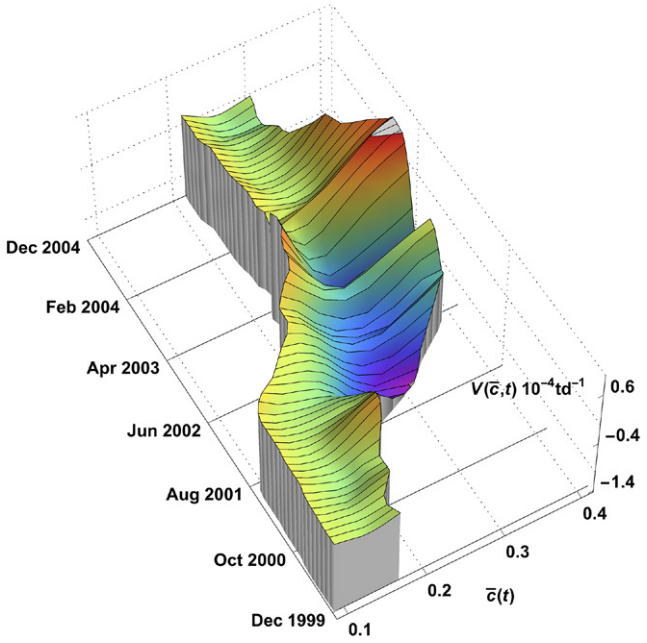

Figure 9. Time evolution of the potential function (22) in the period from early 1999 to early 2006, which covers the dot-com bubble. The estimation is done on a time window of two trading years which is moved in steps of one trading month. The dates mark the time points in the middle of the estimation time windows. Representation is according to equation (30).

hierarchical structure of the states we estimate $V(\bar{c}, t)$ for the points (24). We therefore account for displacements within the classes including state transitions. The resulting potential functions for the three classes are shown in figures 10(a)–(c). These curves envelope the potential functions of the market states of the corresponding class.

Similar to the envelopes we estimated $V(\bar{c})$ on the entire time period at once, as shown in figure 10(d). The potential function of each market state has a distinct position along the first principal component, i.e. a distinct value of $\bar{c}(t)$. We therefore conclude that, while the market is in a given (stable) state, the mean correlation coefficient fluctuates around a mean value, which is defined by the minimum of the potential function, see figures 10 and 6. As we showed in section 2.2, the movement along the first principal component is given by the time evolution of $\bar{c}(t)$. Hence the market dynamics within a fixed state is given by the movement along the second and higher principal components, see figures 3 and 4. Large changes of the mean correlation coefficient yield state transitions. The market is therefore 'hopping' from state to state in the potential landscape, which is shown in figure 10(d). For consistency we calculate the daily steps

$$S(t) = \| \vec{c}(t+1) - \vec{c}(t) \| \tag{31}$$

of the market and the absolute increments

$$\Delta(t) = \left| \bar{c}(t+1) - \bar{c}(t) \right|. \tag{32}$$

of $\bar{c}(t)$. Figures 11(a) and (b) shows the distribution of the steps (31) and the increments (32) within market states compared to the jumps during a state transition. Both

J. Stat. Mech. (2015) P08011

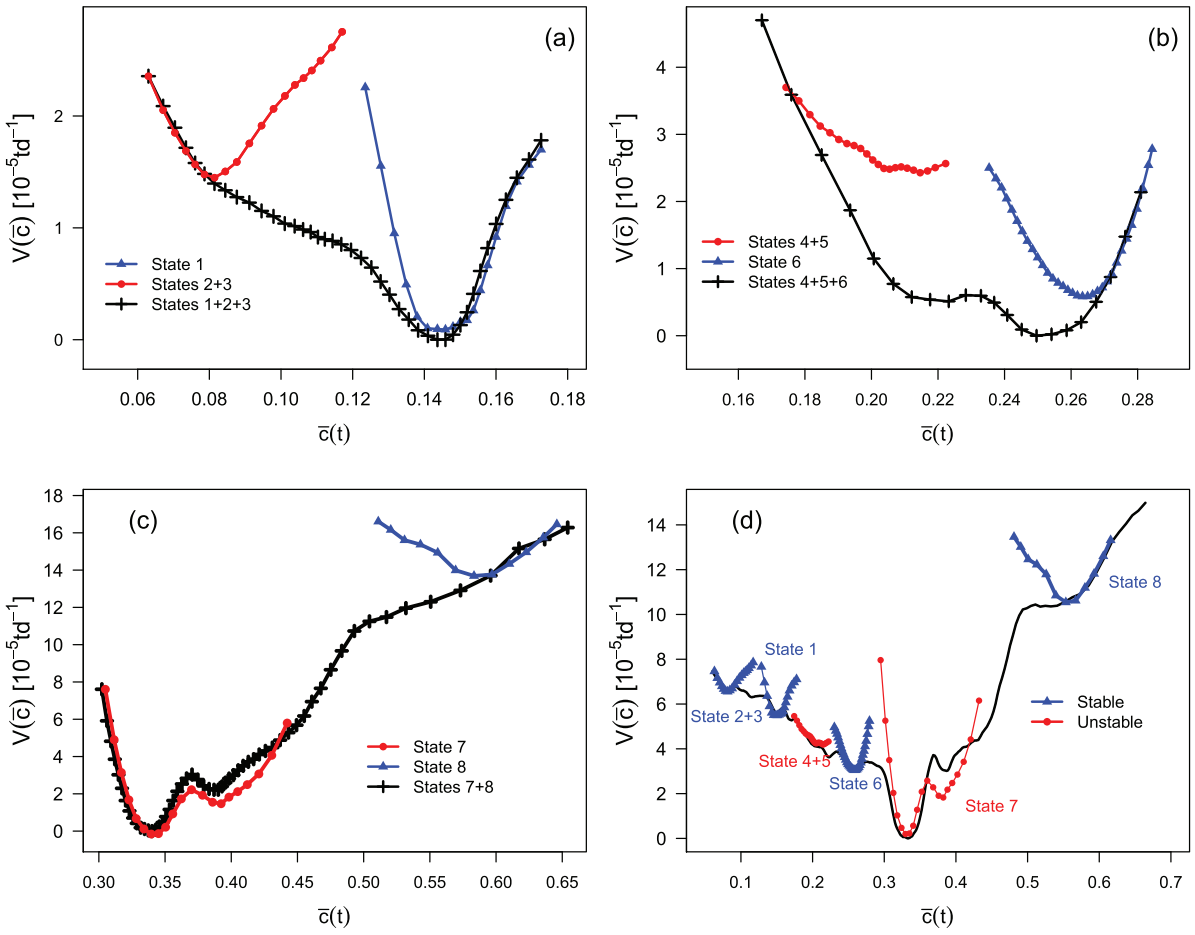

Figure 10. (a)–(c) The potential function (22) of the displacements within the states (23) (filled circles and triangles). The potential function of displacement within the three groups (24) is represented by the envelope line with crosses (+). (d) The overall potential landscape $V(\bar{c})$ estimated on the entire time series $\bar{c}(t)$ at once.

the steps and especially the increments are on average larger during state transitions than within states as we claimed.

6. Conclusion

The combination of geometric data analysis and stochastic methods sheds new light on the collective dynamics of complex systems. We applied these techniques to stock market data and evaluated the correlation structure on a sliding time window for a period of 21 years. The collective market dynamics in terms of the principal components is given by the average correlation coefficient. We extracted the underlying stochastic process which turns out to have a time-independent stochastic term and a time-dependent deterministic term. The latter is represented graphically as a potential landscape and provides information on stability and system fixed points. We established

J. Stat. Mech. (2015) P08011

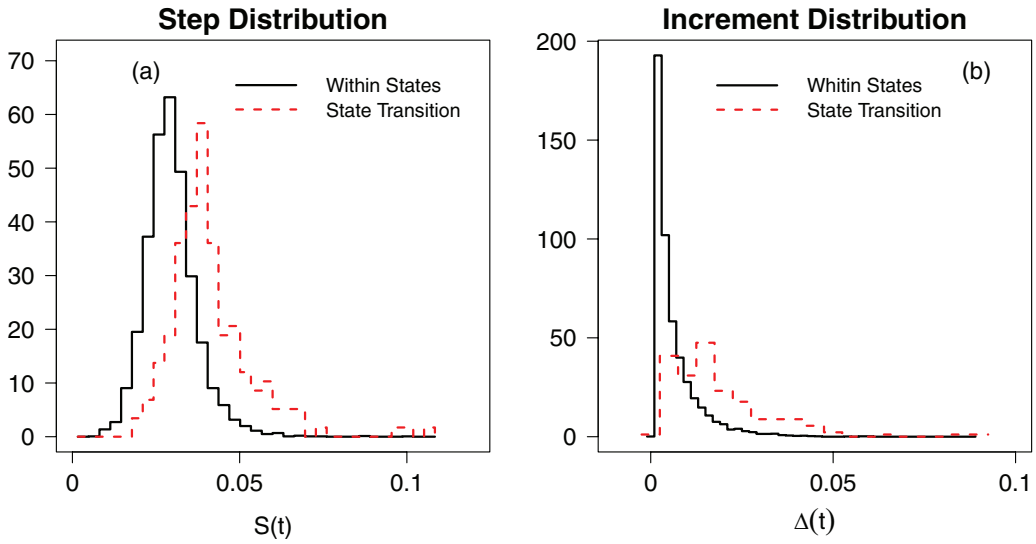

Figure 11. Empirical histograms of (a) the steps (31) and (b) the increments (32) within states (black, solid) and during state transitions (red, dashed).

the connection between distinct historical periods on the market and the time evolution of the potential function. The non-stationary market dynamics can be attributed to changes in the deterministic part of the collective market dynamics. We identified quasi-stationary states of the market following [12] and distinguished three main classes of market dynamics: calm, intermediate and turbulent states. To quantify the market states dynamics we estimated the potential functions, accounting only for displacements within a fixed state. In a given state the average correlation fluctuates around a distinct mean value, which defines a fixed point. The market dynamics within a market state is given by the movement along higher principal components. State transitions are reflected in large changes of the average correlation and correspond to the hopping in the potential landscape. Our results are consistent with the random matrix approach of [42] and contribute to a better overall understanding of market dynamics. While we highlighted the application to financial data in this paper, our approach should prove useful for the study of any quasi-stationary complex system.

Acknowledgment

Yuriy Stepanov acknowledges the Hans-Böckler-Foundation for financial support.

References

[1] Friedrich R, Peinke J, Sahimi M and Tabar M R R 2011 *Phys. Rep.* **506** 87–167
[2] Onnela J P, Chakraborti A, Kaski K, Kertész J and Kanto A 2003 *Phys. Scr.* **106** 48
[3] Bonanno G, Caldarelli G, Lillo F, Micciché S, Vandewalle N and Mantegna R N 2004 *Eur. Phys. J.* B **38** 363–71
[4] Tumminello M, Aste T, Di Matteo T and Mantegna R N 2005 *Proc. Natl Acad. Sci.* **102** 10421–6
[5] Mizuno T, Takayasu H and Takayasu M 2006 *Physica* A **364** 336–42
[6] Tumminello M, Lillo F and Mantegna R N 2010 *J. Econ. Behav. Organ.* **75** 40–58
[7] Musmeci N, Aste T and Di Matteo T 2015 *J. Netw. Theory Finance* **1** 1–22

J. Stat. Mech. (2015) P08011

[8] Mantegna R N 1999 *Eur. Phys. J.* B **11** 193–7
[9] Steinbach M, Karypis G and Kumar V 2000 A comparison of document clustering techniques *Technical Report* 00-034 (Minneapolis, MN: University of Minnesota)
[10] Jain A K 2010 *Pattern Recognit. Lett.* **31** 651–66
[11] Song W M, Di Matteo T and Aste T 2012 *PLoS ONE* **7** e31929
[12] Münnix M C, Shimada T, Schäfer R, Leyvraz F, Seligman T H, Guhr T and Stanley H E 2012 *Sci. Rep.* **2** 644
[13] Pozzi F, Di Matteo T and Aste T 2013 *Sci. Rep.* **3** 1665
[14] Pearson K 1901 *Phil. Mag.* **2** 559–72
[15] Hotelling H 1933 *J. Educ. Psychol.* **24** 417–520
[16] Laloux L, Cizeau P, Potters M and Bouchaud J P 2000 *Int. J. Theor. Appl. Finance* **3** 391–7
[17] Plerou V, Gopikrishnan P, Rosenow B, Amaral L A N, Guhr T and Stanley H E 2002 *Phys. Rev.* E **65** 066126
[18] Lee J A and Verleysen M 2007 *Nonlinear Dimensionality Reduction* (New York: Springer)
[19] Siegert S, Friedrich R and Peinke J 1998 *Phys. Lett.* A **243** 275–80
[20] Friedrich R, Siegert S, Peinke J, Lück S, Siefert M, Lindemann M, Raethjen J, Deuschl G and Pfister G 2000 *Phys. Lett.* A **271** 217–22
[21] Friedrich R, Peinke J and Renner C 2000 *Phys. Rev. Lett.* **84** 5224–7
[22] Renner C, Peinke J and Friedrich R 2001 *Physica* A **298** 499–520
[23] Hutt A, Svensén M, Kruggel F and Friedrich R 2000 *Phys. Rev.* E **61** R4691
[24] Vasconcelos V V, Raischel F, Haase M, Peinke J, Wächter M, Lind P G and Kleinhans D 2011 *Phys. Rev.* E **84** 031103
[25] Rinn P, Stepanov Y, Peinke J, Guhr T and Schäfer R 2015 *Europhys. Lett.* **110** 68003
[26] Kenett D Y, Shapira Y, Madi A, Bransburg-Zabary S, Gur-Gershgoren G and Ben-Jacob E 2010 *AUCO Czech Econ. Rev.* **4** 330–41
[27] Song D M, Tumminello M, Zhou W X and Mantegna R N 2011 *Phys. Rev.* E **84** 026108
[28] Buccheri G, Marmi S and Mantegna R N 2013 *Phys. Rev.* E **88** 012806
[29] Chetalova D, Schmitt T, Schäfer R and Guhr T 2015 *Int. J. Theor. Appl. Finance* **18** 1550012
[30] Schäfer R and Guhr T 2010 *Physica* A **389** 3856–65
[31] Sussmann H J 1978 *Ann. Probab.* **6** 19–41
[32] van Kampen N 1981 *J. Stat. Phys.* **24** 175–87
[33] Dunkel J 2006 Chapter: Langevin–Gleichungen mit nichtlinearer Reibung *Irreversible Prozesse und Selbstorganisation* ed H M Schimansky-Geier and T Pöschel (Berlin: Logos) pp 11–21
[34] Itô K 1944 *Proc. Imperial Acad.* **20** 519–24
[35] Risken H 1996 *The Fokker–Planck Equation: Methods of Solution and Applications* (Lecture Notes in Mathematics) (Berlin: Springer)
[36] Wosnitza J H and Leker J 2014 *Physica* A **401** 228–50
[37] Ball C A and Torous W N 2000 *J. Empir. Finance* **7** 373–88
[38] Burtschell X, Gregory J and Laurent J P 2005 *J. Credit Risk* **13** 31–62
[39] van Emmerich C 2006 Modelling correlation as a stochastic process *Technical Report* Bergische Universität Wuppertal
[40] Ma J 2009 *Ann. Econ. Finance* **10** 303–27
[41] Ankirchner S and Heyne G 2012 *Finance Stoch.* **16** 17–43
[42] Chetalova D, Schäfer R and Guhr T 2015 *J. Stat. Mech.* P01029

J. Stat. Mech. (2015) P08011