Quantitative Finance, 2016
Vol. 16, No. 11, 1713–1724, http://dx.doi.org/10.1080/14697688.2016.1175656

Routledge
Taylor & Francis Group

Early warning of large volatilities based on recurrence interval analysis in Chinese stock markets

ZHI-QIANG JIANG†‡, ASKERY CANABARRO‡§, BORIS PODOBNIK¶‖,
H. EUGENE STANLEY‡ and WEI-XING ZHOU*†

†School of Business and Research Center for Econophysics, East China University of Science and Technology, Shanghai 200237, China
‡Department of Physics and Center for Polymer Studies, Boston University, Boston, MA 02215, USA
§Grupo de Física da Matéria Condensada, Núcleo de Ciências Exatas, Universidade Federal de Alagoas, Campus Arapiraca, 57309-005 Arapiraca, AL, Brazil
¶Faculty of School Engineering, University of Rijeka, 51000 Rijeka, Croatia
‖Zagreb School of Economics and Management, 10000 Zagreb, Croatia

(*Received 1 August 2015; accepted 25 March 2016; published online 24 May 2016*)

Forecasting extreme volatility is a central issue in financial risk management. We present a large volatility predicting method based on the distribution of recurrence intervals between successive volatilities exceeding a certain threshold Q, which has a one-to-one correspondence with the expected recurrence time τ_Q. We find that the recurrence intervals with large τ_Q are well approximated by the stretched exponential distribution for all stocks. Thus, an analytical formula for determining the hazard probability $W(\Delta t|t)$ that a volatility above Q will occur within a short interval Δt if the last volatility exceeding Q happened t periods ago can be directly derived from the stretched exponential distribution, which is found to be in good agreement with the empirical hazard probability from real stock data. Using these results, we adopt a decision-making algorithm for triggering the alarm of the occurrence of the next volatility above Q based on the hazard probability. Using the 'receiver operator characteristic' analysis, we find that this prediction method efficiently forecasts the occurrence of large volatility events in real stock data. Our analysis may help us better understand reoccurring large volatilities and quantify more accurately financial risks in stock markets.

Keywords: Extreme volatility; Risk estimation; Recurrence interval; Large volatility forecasting; Distribution; Hazard probability

JEL Classification: G130, G110

1. Introduction

Predicting extreme volatility events in financial markets is essential in risk estimation. A standard approach to extreme event prediction is to find the precursory patterns prior to an extreme event or to quantify the probability that a given pattern is a precursor to an extreme event (Hallerberg *et al.* 2007, Hallerberg and Kantz 2008). Bogachev and Bunde (2009a, 2011) propose a new method based on the statistics of the recurrence intervals between events exceeding a threshold to determine the risk probability $W(\Delta t|t)$ that an extreme event will occur within the next Δt intervals when the last extreme event occurred t periods ago. They find that when examining real market data and model data with a low level of noise, the predicting method based on recurrence interval analysis produces better forecasts than the method based on precursor pattern recognition (Bogachev and Bunde 2009a, 2011).

Understanding the recurrence interval, defined as the waiting time between consecutive events with values greater than a predefined threshold Q, is essential in uncovering the underlying laws governing extreme events in many fields. Recurrence interval analysis has been carried out on many kinds of time series in predicting the probability that an extreme event will occur, including records of climate (Bunde *et al.* 2004, 2005), seismic activities (Corral 2003, Saichev and Sornette 2006), energy dissipation rates of three-dimensional turbulence (Liu *et al.* 2009), heartbeat intervals in medical science (Bogachev *et al.* 2009), precipitation and river run-off (Bogachev and Bunde 2012), Internet traffic (Bogachev and Bunde 2009b, Cai *et al.* 2009), financial volatilities

*Corresponding author. Email: wxzhou@ecust.edu.cn

(Yamasaki *et al.* 2005, Xie *et al.* 2014), equity returns (Yamasaki *et al.* 2006, Bogachev *et al.* 2007, Bogachev and Bunde 2008, 2009a, Ren and Zhou 2010a, Ludescher *et al.* 2011, He and Chen 2011, Meng *et al.* 2012, Suo *et al.* 2015), and trading volumes (Podobnik *et al.* 2009, Ren and Zhou 2010b, Li *et al.* 2011). An improved method for estimating the value at risk (VaR) in financial markets has been proposed based on the recurrence interval between the last two returns below $-Q$. This method is significantly more accurate than traditional estimates based on the overall or local return distributions (Bogachev and Bunde 2009a, Ludescher *et al.* 2011).

To accurately estimate the risk probability and the VaR based on recurrence interval analysis, we need the distribution and memory behaviour of a set of recurrence intervals between extreme events. It is found that the recurrence intervals of the time series in many different fields exhibit fat-tailed distributions and long and short range memories, indicating that extreme events cannot be described by the Poisson process. Unlike long and short range memory behaviours, which are easily testable using conditional distribution analysis and the DFA method, the distribution form of recurrence intervals is still elusive. For example, in financial markets, the recurrence intervals of different data types (return, volatility and trading volume), different data resolutions (minute-by-minute and daily) and different markets fit different distributions, including the power-law, stretched exponential and q-exponential. It has been found that the recurrence interval distribution above a fixed threshold has a power-law tail for the daily volatilities in the Japanese market (Kaizoji and Kaizoji 2004, Yamasaki *et al.* 2005), the minute-by-minute volatilities in the Korean market (Lee *et al.* 2006) and the Italian market (Greco *et al.* 2008), the daily returns in the US stock markets (Bogachev *et al.* 2007, Bogachev and Bunde 2008, 2009a), the minute-by-minute returns in the Chinese markets (Ren and Zhou 2010a), and the minute-by-minute trading volume in the US markets (Li *et al.* 2011) and the Chinese markets (Ren and Zhou 2010b). A number of studies ranging from daily to high-frequency data and from developed to emerging markets (Wang *et al.* 2006, 2007, Jung *et al.* 2008, Qiu *et al.* 2008, Ren *et al.* 2009a, 2009b, Jeon *et al.* 2010, Wang and Wang 2012, Xie *et al.* 2014), have reported that the distribution of the recurrence intervals of financial volatility is a stretched exponential. The stretched exponential distribution is also observed in the recurrence time between returns above a given positive threshold or below a negative threshold for the index spot and futures in the Chinese future markets (Suo *et al.* 2015), which is in contrast to the q-exponential distribution observed in the recurrence intervals between the losses in financial returns (Ludescher *et al.* 2011, Ludescher and Bunde 2014).

In this paper, we describe the materials and methods in section 2, determine the distribution of the recurrence intervals between large volatilities in section 3 and report the hazard probability results and predicting algorithm performance in section 4. In section 5, we summarize our findings.

2. Materials and methods

2.1. Data description

To carry out a detailed recurrence interval analysis of the Chinese stock markets, we include as many Chinese stocks in our analysing sample as possible. The minute-by-minute price data of all stocks in the Chinese markets are extracted from the RESSET financial database. The extracting period is from 26 July 1999 to 30 December 2011, which is the maximum spanning period allowed in the RESSET database. To ensure that the recurrent interval results between the top 1% volatilities will have more than 1000 data points, we select only those stocks that have a minimum of two years of trading records. The large sample size lowers the error rate when we use a maximum likelihood estimation (MLE) to fit the distributions. Finally, we have 1820 stocks in our sample, including 799 A-shares, 54 B-shares and 51 ChiNext shares in the Shenzhen market, and 862 A-shares and 54 B-shares in the Shanghai market.

2.2. Definition of volatilities and recurrence intervals

Following the review of volatility estimator in Bollen and Inder (2002), for a given minute-by-minute price series $I(t)$, the minute-by-minute volatility $\omega(t)$ can be estimated using

$$\omega(t) = |\ln I(t) - \ln I(t-1)|. \tag{1}$$

Such definition is also widely used in the reference on recurrence interval analysis of financial volatilities (Yamasaki *et al.* 2005, Wang *et al.* 2006, 2008, 2009, Xie *et al.* 2014). Using absolute return as financial volatility not only has the advantage of considering the extreme negative and positive returns at the same time, but also allows us to directly compare our results with the results reported in the literature.

In order to eliminate the influence of the daily periodic patterns, we remove the intraday patterns from the volatility series $\omega(t)$ on each trading day,

$$\omega'(s) = \omega(s)/G(s), \tag{2}$$

where $G(s) = \sum_i^N \omega(i,s)/N$. Here, $\omega(i,s)$ represents the volatility at time s on day i. The normalized volatility series $v(t)$ is then obtained by dividing $\omega'(t)$ by its standard deviation,

$$v(t) = \frac{\omega'(t)}{\sqrt{[\langle \omega'(t)^2\rangle - \langle \omega'(t)\rangle^2]}}, \tag{3}$$

where $\langle z \rangle$ means the average value of z.

The focus of our study is the recurrence interval between the normalized volatilities exceeding a predefined threshold Q. To compare the results between different stocks, we quantify Q by its mean recurrence time τ_Q. There is a one-to-one correspondence between Q and τ_Q, such that (Podobnik *et al.* 2009)

$$\frac{1}{\tau_Q} = \int_Q^\infty p_v(v)\mathrm{d}v, \tag{4}$$

where $p_v(v)$ is the probability distribution of the volatility. Using copula, Chicheportiche and Chakraborti (2014) also prove this equation and state that this equation is universal no matter what the underlying dependence structure in the analysing series is. Here, we restrict τ_Q to a range of [20, 100]. This range corresponds to the extreme volatilities from a top value of 5–1%, which is often considered in risk estimation.

3. Distribution of recurrence intervals between large volatilities

The distributions of recurrence intervals are found to exhibit a scaling behaviour for the large volatilities filtered by different thresholds (Yamasaki *et al*. 2005, Wang *et al*. 2006), which means that the interval distribution is independent of the thresholds and thus the distribution of recurrence intervals between extreme events can be inferred from the distribution of recurrence intervals between non-extreme events. However, such scaling behaviours are rejected by follow-up analysis (Wang *et al*. 2008, 2009, Ren and Zhou 2008), which states that the extreme event filtering threshold should have an influence on the recurrence interval distribution. This behaviour is corroborated by the fact that the estimated distributional parameters are found to have a strong dependence on the thresholds when the recurrence intervals are fitted by given distribution functions, such as stretched exponential distribution (Xie *et al*. 2014, Suo *et al*. 2015) and q-exponential distribution (Ludescher *et al*. 2011, Chicheportiche and Chakraborti 2014). More interestingly, Ludescher *et al*. (2011) and Ludescher and Bunde (2014) argue that the distribution of recurrence intervals depends only on the mean recurrence interval τ_Q, and not on a specific asset or on the time resolution of the data. According to their analysis, the distribution of recurrence intervals from different financial series should overlap on the same curve for the same threshold. The above discussions motivate us to understand the following three problems for the recurrence interval between large volatilities in the Chinese stock markets: (1) whether there is a scaling behaviour for the recurrence intervals between the large volatilities with different thresholds for individual stocks; (2) whether the recurrence intervals between large volatilities with the same threshold exhibit the same distribution for different stocks; and (3) whether the market dynamics have any influence on the distribution of recurrence intervals.

In order to have an overview of the distribution of the recurrence intervals between large volatilities, we firstly plot the recurrence interval distribution of randomly chosen stocks in log–log scale in figure 1. As shown in figure 1(a), the distributions of the recurrence interval between large volatilities obtained from different thresholds are scaled by their τ_Q. The distribution of different stocks is shifted vertically for better visibility. One can see that the scaled recurrence interval distributions corresponding to different thresholds are overlapping on the same curve for the five stocks. Obviously, the appearance of the collapsing behaviours indicates that the recurrence interval distributions possess scaling behaviours for the five randomly chosen stocks. Figure 1(b) illustrates the collapsing behaviours of the recurrence interval distributions obtained from the same τ_Q for different stocks. Ten stocks are randomly chosen to present the distribution overlapping behaviour. In order to increase visibility, the distribution curves of different τ_Q values are transformed by a factor. It is found that for the same τ_Q value, the recurrence time distributions of different stocks are overlapping on the same curve. This means that the return intervals between volatilities that exceed a threshold Q may exhibit a universal distribution for different stocks when τ_Q is fixed. The results in figure 1 provide an amazing picture that the scaled recurrence interval distribution is independent of the τ_Q for individual stocks and the recurrence intervals with the same τ_Q have the same distribution across different stocks, which leads to the hypothesis that the distribution of scaled recurrence interval between large volatilities is universal, depends neither on the value of τ_Q nor on specific stocks.

To quantitatively test the hypothesis, the best strategy is to find a suitable distribution function to fit the recurrence interval distribution. This allows us to verify the above hypothesis by comparing with the estimated distributional parameters corresponding to different stocks and thresholds. It is well known that the memory behaviour in the underlying process plays a very important role in the distribution form of the recurrence intervals. The memory-less process always results in an exponential distribution of recurrence time. For the process with long memory, the situation is much more complicated. If only linear long memory is incorporated, the stretched exponential and Weibull distribution of recurrence intervals are supported both in theoretical derivations (Santhanam and Kantz 2008, Olla 2007) and numerical experiments (Altmann and Kantz 2005, Pennetta 2006, Eichner *et al*. 2007). Furthermore, power-law distributions of recurrence time are uncovered in synthetic processes with non-linear long memory behaviours (multifractal processes) (Bogachev *et al*. 2007, 2008a, Bogachev and Bunde 2008). As with the synthetic processes, the recurrence interval analysis on the realistic processes, for example the empirical analysis in financial markets, also provides inconsistent results, such that the recurrence time is found to follow a stretched exponential distribution (Wang *et al*. 2006, 2007, Jung *et al*. 2008, Qiu *et al*. 2008, Jeon *et al*. 2010, Wang and Wang 2012, Suo *et al*. 2015), a power-law distribution with an exponential cut-off (Kaizoji and Kaizoji 2004, Yamasaki *et al*. 2005, Lee *et al*. 2006, Greco *et al*. 2008), and a q-exponential distribution (Ludescher *et al*. 2011, Ludescher and Bunde 2014). The q-exponential distribution is also reported to nicely fit the inter-trade time, which is the waiting time between consecutive trades (Politi and Scalas 2008, Scalas *et al*. 2004, Jiang *et al*. 2008). Here, we propose to use four candidate distributions to fit the recurrence intervals between large volatilities in the Chinese stock markets. Our purpose is to find the best fits out of the four distributions for the following predicting analysis. The following are four candidate distributions: the stretched exponential distribution,

$$p_s(\tau) = a \exp[-(b\tau)^{\mu}], \tag{5}$$

the power law distribution with an exponential cut-off,

$$p_p(\tau) = c\tau^{-\gamma-1} \exp(-k\tau), \tag{6}$$

the q-exponential distribution,

$$p_q(\tau) = (2-q)\lambda[1+(q-1)\lambda\tau]^{-\frac{1}{q-1}}, \tag{7}$$

and the Weibull distribution,

$$p_w(\tau) = \frac{\zeta}{d}\left(\frac{\tau}{d}\right)^{\zeta-1} \exp\left[-\left(\frac{\tau}{d}\right)^{\zeta}\right]. \tag{8}$$

We have the following two equations $\int_0^{\infty} p(\tau)\mathrm{d}\tau = 1$ and $\int_0^{\infty} \tau p(\tau)\mathrm{d}\tau = \tau_Q$ for any distribution formula of $p(\tau)$, which allow us to reduce the number of estimated parameters for the four candidate distributions. The second equation is universal, independent of any memory structure of the underlying process (Chicheportiche and Chakraborti 2014, 2013). With the two equations, we have $a = \frac{\mu\Gamma(2/\mu)}{\Gamma(1/\mu)^2\tau_Q}$ and $b = \frac{\Gamma(2/\mu)}{\Gamma(1/\mu)\tau_Q}$

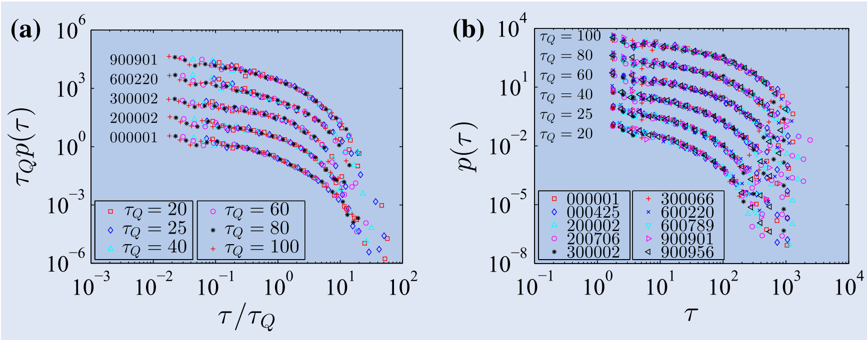

Figure 1. Collapsing behaviours of the recurrence interval distributions. (a) Collapsing behaviours of the distributions of the scaled recurrence intervals for five randomly chosen stocks. The distribution of stocks 200002, 300002, 600220 and 900901 is shifted vertically by a factor of 10, 100, 1000 and 10 000, respectively. (b) Collapsing behaviours of the distributions of recurrence intervals for 10 randomly chosen stocks with the same τ_Q. The distribution curve of $\tau_Q = 25$, 40, 60, 80 and 100 is shifted vertically by a factor of 10, 100, 1000, 10 000 and 100 000, respectively.

for the stretched exponential distribution, $k = \frac{-\gamma}{\tau_Q}$ and $b = (\frac{-\gamma}{\tau_Q})^{-\gamma}\frac{1}{\Gamma(-\gamma)}$ for the power-law distribution with an exponential cut-off, $\lambda = \frac{1}{\tau_Q(3-2q)}$ for the q-exponential distribution and $d = \frac{\tau_Q}{\Gamma(1+1/\zeta)}$ for the Weibull distribution. Obviously, we have only one parameter to estimate for the four candidate distributions. We use the MLE method to estimate the distributional parameters. Because it is very hard to obtain the equation analytically by taking a derivative of the logarithmic likelihood function with the corresponding estimated parameters, we discretize the estimating parameters in their fitting range with a step of 10^{-6} ($\mu \in (0, 1)$, $\gamma \in (-1, 0)$, $q \in (1, 1.5)$, and $\zeta \in (0, 1)$) and evaluate the logarithmic likelihood function $\ln L$ of these discrete values. The discrete value associated with the maximum $\ln L$ is took as the final solution for the estimated parameter.

Figure 2 shows the empirical distributions and the fitting results of the four candidate distributions of the recurrence intervals for two stocks, 000001 and 900956. One can see that in the central regions of the distributions, all four candidates agree with the empirical data. The q-exponential distribution better fits the distribution for small τ_Q and the stretched exponential distribution gives better fits in the tail for large τ_Q. To determine which distribution has the best performance, we utilize likelihoods and KS statistics to assess the validity of the estimates. As we known, KS statistic measures the distance between the fitting cumulative distribution and the empirical cumulative distribution. Here, we have four candidate distributions to fit the recurrence intervals. If one distribution gives the minimum KS statistic, this distribution can be regarded as the closest to the empirical distribution, which indicates the best fit. This is applied in finding the best truncated boundary when fitting to the left-truncated distributions (Clauset *et al.* 2009, Jiang *et al.* 2013). Therefore, the candidate distribution giving the maximum likelihood or the minimum KS statistic is considered to fit the recurrence intervals best.

Taking stock 000001 as an example, figure 3(a) highlights the candidate distribution which has the maximum likelihood when it is fitted to the recurrence intervals for each τ_Q. Figure 3(b) is the same as (a), except that we highlight the distributions on the basis of the minimum KS statistics. The symbols 's', 'p', 'q' and 'w' correspond to stretched exponential distribution, power-law distribution with an exponential cut-off, q-exponential distribution and Weibull distribution, respectively. Both panels indicate that the q-exponential distribution gives the best fits when τ_Q is not large and the stretched exponential distribution performs the best for large values of τ_Q. However, the transition point is different for the two validity measurements. The transition point from q-exponential distribution to stretched exponential distribution is close to $\tau_Q = 90$ for KS statistics, which is much larger than that for likelihoods, $\tau_Q = 60$. Such results allow us to further classify the stocks into one group for each τ_Q if the same distribution fits their recurrence intervals best.

We count the number of stocks in each group for each τ_Q and each validity measurement and report the results in table 1. One can observe that more than 90% of stocks are in q-exponential and stretched exponential groups for all values of τ_Q and both validity measurements favour the q-exponential distribution (respectively, stretched exponential distribution) when τ_Q approaches 20 (100). However, the classification of stocks based on likelihoods and KS statistics does not tell us anything about the significance of the fits. Following the methods (Jiang *et al.* 2013), we employ both KS tests and CvM tests to check the goodness of the fits in each group for each τ_Q. The null hypothesis H_0^x for both tests is that the data are drawn from a testing distribution. The testing distribution for a given group corresponds to the distribution that fits the recurrence intervals best for the stocks in that group. The stocks whose recurrence intervals cannot reject the hypothesis H_0^x at the significant level of 0.01 for either of the two tests are labelled as passing the statistical tests.

We also list the number of stocks passing the statistical tests in table 1. We find that only a small number of stocks are found to be in the group of power-law distribution with an exponential cut-off and Weibull distribution. However, most of them do not fail the statistical tests. It is also observed that only a small number of stocks in q-exponential group pass the statistical tests for all τ_Q. This means that q-exponential distribution is not suitable to fit the recurrence intervals between extreme volatilities in the Chinese stock markets, although the validity measurements favour it for small τ_Q. In contrast, the number of stocks passing the tests in stretched exponential group keeps increasing with the increment of τ_Q. The percentage of stocks out of the Chinese markets whose recurrence intervals cannot reject the stretched exponential distribution is more than 60%

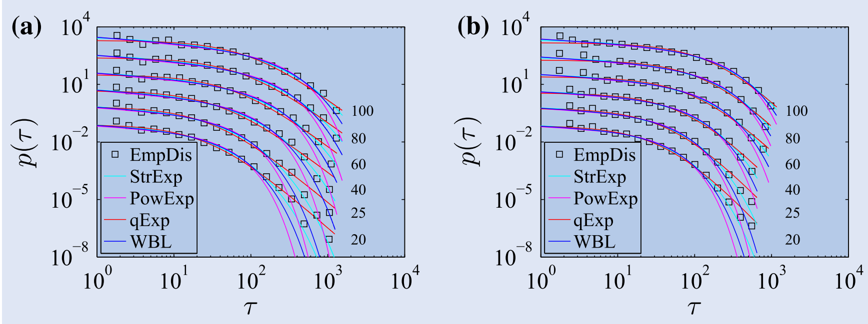

Figure 2. Plots of distribution fits to recurrence intervals for two stocks. The open markers are the empirical distributions and the solid lines are the fits to the four candidate distributions. For better visibility, the curves of $\tau_Q = 25$, 40, 60, 80 and 100 are shifted vertically by a factor of 10, 100, 1000, 10 000 and 100 000, respectively. (a) Stock 000001. (b) Stock 900956.

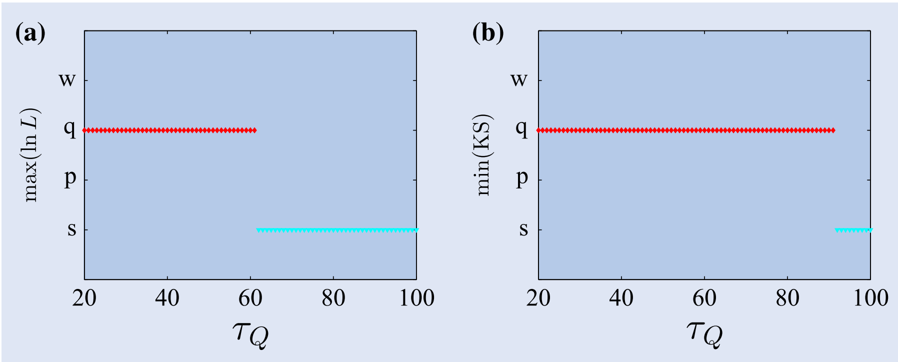

Figure 3. Validity of the fits to the four candidate distributions. (a) Assessing the validity of the candidate distributions with the likelihoods for stock 0000001. The symbols ‘s’, ‘p’, ‘q’ and ‘w’ represent stretched exponential distribution, power-law distribution with an exponential cut-off, q-exponential distribution and Weibull distribution, respectively. The distribution that has the maximum likelihood is highlighted by a marker for each τ_Q. (b) The same as (a), but the distribution is highlighted based on the minimum KS statistics.

(respectively, 56%) for the validity measurements of likelihoods (respectively, KS statistics) when $\tau_Q = 100$. The results indicate that the stretched exponential distribution is a good candidate to fit the recurrence intervals with large τ_Q in the Chinese stock markets. Another intriguing observation in table 1 is that very few stocks pass the statistical tests for small τ_Q. It could be attributed to that the parameter estimating methods are based on the continuous distribution formula. Low τ_Q will lead to a sample of recurrence intervals containing many small discrete values, which makes the estimator biased.

Despite the statistical tests for goodness of fits do not give affirmative results of accepting the null hypothesis H_0^x for all distributional fits, the four candidate distributions are still a very good approximation to the empirical recurrence interval distributions, as evidenced in figure 2. Therefore, it is feasible to discuss the scaling behaviour for individual stocks with resort to the dependence between the distribution parameters and τ_Q. As stated above, if the recurrence intervals exhibit a scaling behaviour, the estimated parameters of the four distributions should be independent of τ_Q, which means that the slope should be zero if the estimated parameters are regressed against τ_Q. We plot the estimate parameters as a function of τ_Q in figure 4 for stock 000001 (a) and stock 900956 (b). It is observed that the four parameters present an asymptotic horizontal line with respect to τ_Q for both stocks. The least linear square fits of μ, γ, q and ζ with respect to τ_Q give the four slopes of $\kappa_\mu = -8.84 \times 10^{-4}$, $\kappa_\gamma = 1.1 \times 10^{-3}$, $\kappa_q = 3.43 \times 10^{-4}$, and $\kappa_\zeta = -6.79 \times 10^{-4}$ (respectively, -8.87×10^{-4}, 1.2×10^{-3}, 1.94×10^{-4} and -6.92×10^{-4}) for stock 000001 (900956). One can observe that the slopes are small but not vanishing, indicating the existence of a weak trend between the distribution parameters and τ_Q. Taking the stretched exponential distribution as an example, the parameter μ will at least decrease 0.07 when τ_Q varies from 20 to 100. Such weak trend is a sign of no scaling behaviour in recurrence intervals for both stocks.

We further propose a strict statistical test in the spirit of bootstrapping to check whether there exists a scaling behaviour in recurrence intervals for individual stocks. The null hypothesis H_0^y is that the distribution parameters (μ, γ, q and ζ) are independent of τ_Q. If the null hypothesis H_0^y cannot be rejected, the recurrence intervals can be argued to possess a scaling behaviour for a given stock. For a given distribution, we estimate the slope κ by linearly regressing its parameter against τ_Q. We define the statistics as the slope κ. In order to accumulate an ensemble of such statistic, which allows us to generate a distribution, the parameter of the given distribution is reshuffled for 1000 times and the slope κ^s between shuffled data and τ_Q is determined. The shuffled procedure can destroy any dependence between the distribution parameter and τ_Q, which gives a quite suitable null model to assess the significance of the statistical test. For a given distribution, we perform the statistical test on each stock and find that 1586 stocks reject the hypothesis H_0^y for all four fitting distributions at the significant level of 0.01. This implies that these stocks do not exhibit a scaling behaviour in their recurrence intervals.

Table 1. Number of stocks in each distribution group. For each value of τ_Q, the stocks are assigned into a given distribution group if the fitting of their recurrence intervals to that distribution gives the maximum likelihood or the minimum KS statistic. Statistical tests (KS tests and CvM tests) are performed to check the goodness of fits for each stock in each group. The stocks whose recurrence intervals cannot reject the hypothesis H_0^x at the significant level of 0.01 for either of the two tests are labelled as passing the statistical tests. 'total' means the number of total stocks and 'pass' stands for the number of stocks passing the statistical tests.

	Likelihood								KS statistic							
	StrExp		PowExp		qExp		WBL		StrExp		PowExp		qExp		WBL	
τ_Q	Total	Pass	Total	Pass	Total	Pass	Total	Pass	Total	Pass	Total	Pass	Total	Pass	Total	Pass
20	59	0	0	0	1761	0	0	0	0	0	93	0	1727	0	0	0
25	322	3	0	0	1498	29	0	0	0	0	15	0	1805	18	0	0
30	710	39	0	0	1110	46	0	0	0	0	0	0	1820	52	0	0
35	1040	80	0	0	780	34	0	0	5	0	0	0	1815	60	0	0
40	1266	128	0	0	554	42	0	0	22	9	1	1	1796	92	1	0
45	1427	189	0	0	393	47	0	0	49	28	2	1	1763	108	6	4
50	1525	278	0	0	293	52	2	1	104	54	5	4	1702	112	9	6
55	1581	376	2	2	234	57	3	2	217	72	9	8	1575	99	19	15
60	1628	510	2	2	182	50	8	4	413	94	14	13	1363	97	30	24
65	1658	647	3	3	144	52	15	7	684	183	19	16	1079	93	38	34
70	1662	743	7	7	131	52	20	15	936	337	23	21	817	94	44	39
75	1666	839	8	8	115	48	31	27	1173	524	28	27	564	89	55	50
80	1673	915	14	14	95	41	38	33	1345	693	40	38	381	83	54	49
85	1662	985	20	19	91	39	47	42	1427	824	45	43	281	80	67	62
90	1663	1036	28	27	78	33	51	46	1481	910	53	50	211	69	75	63
95	1650	1060	36	35	67	26	67	57	1535	982	56	53	146	55	83	68
100	1642	1090	42	41	59	24	77	63	1542	1028	67	64	118	50	93	71

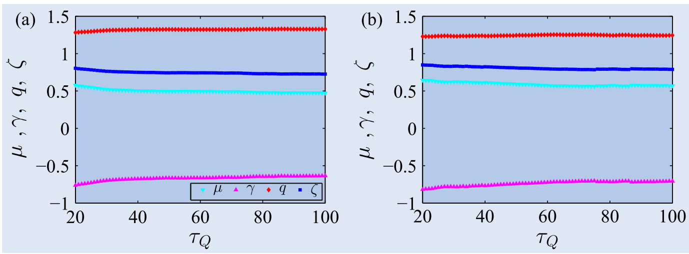

Figure 4. Plots of the estimating parameters with respect to τ_Q for the stretched exponential distribution (μ), the power-law distribution with exponential cut-off, the q-exponential distribution (q) and the Weibull distribution (ζ). (a) Stock 000001. (b) Stock 900956.

For the same τ_Q and a given distribution, we regard the corresponding estimated parameters of different stocks as one sample and estimate the mean, standard deviation, skewness and kurtosis. The results are reported in table 2. It is observed that the mean value of the distribution parameters exhibit a weak trend against τ_Q, which is in agreement with the above analysis on individual stocks. Another intriguing finding is that the standard deviation is stable with the increment of τ_Q for each candidate distribution. Such narrow standard deviation implies that the recurrence intervals of different stocks roughly conform to the same distribution when τ_Q has the same value (Ludescher *et al*. 2011, Ludescher and Bunde 2014). We also observe that the skewness does not equal to 0 and the kurtosis is greater than 3, which indicate that the distributions of estimated parameters are non-normal. More interesting, the skewness of μ and ζ decreases from positive to negative with the increase in τ_Q, which indicates that the distribution shifts from right-skewed to left-skewed, while for the skewness of γ and q, the situation is the opposite.

In order to determine whether the distribution parameters are influenced by market states, i.e. bull or bear, we use a moving window analysis to track the evolution of estimating parameters (μ, γ, q and β). We fix the window size at 48 months and exclude stocks with trading periods shorter than 97 months. We are left with 948 stocks as subject for our rolling window analysis. The window (ending time) varies from 2003 to 2011 with a step of one year. We also discard first-month trading data for the remaining because first-month records tend to be partial (i.e. do not span an entire month) and the volatilities for new IPOs are excessively large (Guo *et al*. 2013), which is caused by the existence of information asymmetry for IPO firms and the incomplete knowledge about IPO firms for investors (Hussein and Zhou 2014).

For each stock, we perform the same analysis in each window as for in the entire period. The recurrence intervals associated with different values of τ_Q are fitted by the four candidate distributions, which give the corresponding distribution parameters (μ, γ, q and ζ). As stated above, the distribution parameters are universal across different stocks when τ_Q has the same value. We estimate the average and standard deviations of the estimated parameters of different stocks for the same τ_Q in each window. The results are listed in table 3. Generally speaking, the distribution parameters estimated from the moving windows and the entire period share the same

Table 2. The basic statistics of the distribution parameters. μ, γ, q and ζ correspond to the parameters of stretched exponential distribution, power-law distribution with an exponential cut-off, q-exponential distribution and Weibull distribution. We list the mean, standard deviation, skewness and kurtosis for the sample of the parameters from different stocks with the same τ_Q.

	μ				γ				q				ζ			
τ_Q	Mean	Std	Skew	Kurt	Mean	Std	Skew	Kurt	Mean	Std	Skew	Kurt	Mean	Std	Skew	Kurt
20	0.62	0.08	−0.25	3.95	−0.80	0.08	0.66	4.57	1.25	0.06	0.07	3.69	0.83	0.06	−0.77	4.87
25	0.60	0.08	−0.13	4.26	−0.77	0.08	0.61	4.91	1.26	0.06	−0.08	3.95	0.82	0.05	−0.73	5.23
30	0.59	0.08	−0.01	4.45	−0.75	0.07	0.55	5.11	1.26	0.06	−0.24	4.12	0.81	0.05	−0.67	5.45
35	0.58	0.08	0.10	4.51	−0.74	0.07	0.46	5.16	1.27	0.06	−0.37	4.17	0.80	0.05	−0.58	5.50
40	0.57	0.08	0.22	4.58	−0.73	0.07	0.38	5.15	1.27	0.06	−0.49	4.27	0.79	0.05	−0.50	5.50
45	0.56	0.08	0.34	4.66	−0.72	0.07	0.29	5.12	1.28	0.06	−0.61	4.41	0.79	0.05	−0.41	5.46
50	0.55	0.08	0.46	4.81	−0.71	0.07	0.20	5.13	1.28	0.06	−0.72	4.59	0.78	0.05	−0.32	5.46
55	0.55	0.08	0.55	4.94	−0.70	0.07	0.12	5.12	1.28	0.06	−0.81	4.70	0.78	0.05	−0.25	5.44
60	0.54	0.08	0.64	5.06	−0.70	0.07	0.06	5.09	1.28	0.06	−0.88	4.80	0.77	0.05	−0.18	5.41
65	0.54	0.08	0.70	5.11	−0.69	0.07	0.00	5.09	1.28	0.06	−0.93	4.84	0.77	0.06	−0.13	5.39
70	0.53	0.09	0.77	5.18	−0.69	0.07	−0.06	5.05	1.28	0.06	−0.98	4.91	0.77	0.06	−0.08	5.35
75	0.53	0.09	0.78	5.14	−0.68	0.07	−0.07	5.00	1.28	0.06	−1.00	4.89	0.76	0.06	−0.07	5.30
80	0.53	0.09	0.86	5.34	−0.68	0.07	−0.13	5.05	1.28	0.06	−1.06	5.04	0.76	0.06	−0.01	5.32
85	0.53	0.09	0.90	5.39	−0.68	0.07	−0.16	5.03	1.29	0.06	−1.08	5.06	0.76	0.06	0.02	5.29
90	0.52	0.09	0.94	5.44	−0.67	0.07	−0.19	5.00	1.29	0.06	−1.12	5.12	0.76	0.06	0.05	5.25
95	0.52	0.09	0.95	5.44	−0.67	0.07	−0.20	4.94	1.29	0.06	−1.13	5.12	0.76	0.06	0.07	5.19
100	0.52	0.09	0.99	5.47	−0.67	0.07	−0.23	4.89	1.29	0.07	−1.17	5.22	0.76	0.06	0.10	5.14

pattern with respect to τ_Q, corresponding to the decreasing trends for μ and ζ and the increasing trends for γ and q. For the same τ_Q, we again observe that with the evolving of time, μ and ζ decrease and γ and q increase. Such results can be linked to the inefficiency of the Chinese stock markets in recent years, that two market crashes are observed in October 2007 and August 2009 (Jiang *et al.* 2010). Market crashes will result in price synchronicity for different stocks (Song *et al.* 2011, Meng *et al.* 2014, Dai *et al.* Forthcoming) and enhance the memory behaviour for individual stocks (Jiang *et al.* 2014). The stretched exponential parameter μ is found to be the correlation exponent of underlying process (Bunde *et al.* 2005, Bogachev *et al.* 2008b). The increase in memory behaviour will lead to a decrease in μ, which is consistent with our results.

4. Predicting large volatilities

We use the hazard probability $W(\Delta t|t)$ to forecast the occurrence of large volatility events. The $W(\Delta t|t)$ is the probability that there will be additional waiting time Δt before another large volatility event occurs when the previous large volatility event occurred t time ago, which can be formulated as (Sornette and Knopoff 1997, Bogachev *et al.* 2007),

$$W(\Delta t|t) = \frac{\int_t^{t+\Delta t} p(t)\mathrm{d}t}{\int_t^{\infty} p(t)\mathrm{d}t}, \tag{9}$$

where $p(t)$ is the probability distribution of the recurrence intervals between extreme events. This probability is the key early warning measurement for the occurrence of extreme volatilities. The early warning is triggered when the probability $W(\Delta t|t)$ is greater than a predefined hazard threshold. We can theoretically derive this hazard probability if we have the distribution of the recurrence intervals between consecutive extreme volatilities.

If we designate the top 1% of volatility values (corresponding to the mean recurrence time $\tau_Q = 100$) to be extreme events, we can estimate the hazard probability $W_q(\Delta t, t)$ in t when fixing Δt. The recurrence intervals between volatilities with $\tau_Q = 100$ are well approximated by the stretched exponential distribution for most stocks, as evidenced in table 1. This allows us to approach the theoretical hazard probability $W_q(\Delta t, t)$ in terms of the stretched exponential distribution. By substituting equation (5) into equation (9), we obtain

$$W_{\mathrm{SE}}(\Delta t|t) = \frac{\frac{b\mu}{a} - \Gamma_l\left(\frac{1}{\mu}, (bt)^{\mu}\right) - \Gamma_u\left(\frac{1}{\mu}, [b(t+\Delta t)]^{\mu}\right)}{\Gamma_u\left(\frac{1}{\mu}, (bt)^{\mu}\right)}, \tag{10}$$

where $\Gamma_l(s, x)$ and $\Gamma_u(s, x)$ are lower and upper incomplete gamma functions and a and b are dependent on μ. On the other hand, the empirical hazard function W_{emp} can be evaluated as follows,

$$W_{\mathrm{emp}}(\Delta t|t) = \frac{\#(t < \tau \leq t + \Delta t)}{\#(\tau > t)}, \tag{11}$$

where the denominator $\#(\tau > t)$ is the number of recurrence intervals whose values are greater than t, and the numerator $\#(t < \tau \leq t + \Delta t)$ is the number of recurrence intervals which locates in the range of $(t, t + \Delta t]$.

Figure 5 shows the hazard probability for two stocks when $\Delta t = 1, 5$ and 10. The solid curves stand for the analytical solution W_{SE} in equation (10) and the markers represent the empirical hazard probability in equation (11). One can see that in each panel, the curve and the markers decrease slowly and are in good agreement. The decreasing trend of $W(\Delta t|t)$ means that the probability of observing another following large volatility decreases with the elapsing t if an extreme volatility occurs. This implies the existence of potential dependent structures in triggering the large volatilities, which is inconsistent with the clustering behaviour in the volatility series. The oscillations of W_{emp} in panels (e) and (f) are attributed to the poor statistics of recurrence time in the sampling interval of $(t, t + \Delta t]$. Besides stock volatilities, the general formula of hazard probability

Table 3. Results of the rolling window analysis. μ, γ, q and ζ correspond to the parameters of stretched exponential distribution, power-law distribution with an exponential cut-off, q-exponential distribution and Weibull distribution. We list the mean for the sample of the parameters from different stocks in the same window with the same τ_Q, as well as the standard deviation in parentheses.

Moving window	τ_Q = 20	30	40	50	60	70	80	90	100
Panel A: Stretched exponential distribution (μ)									
2003	0.70(0.07)	0.66(0.07)	0.64(0.07)	0.62(0.08)	0.62(0.08)	0.61(0.09)	0.61(0.09)	0.61(0.09)	0.62(0.10)
2004	0.73(0.09)	0.70(0.08)	0.69(0.08)	0.69(0.09)	0.69(0.09)	0.69(0.09)	0.69(0.10)	0.69(0.10)	0.70(0.10)
2005	0.64(0.11)	0.62(0.11)	0.61(0.11)	0.61(0.11)	0.61(0.12)	0.61(0.12)	0.61(0.12)	0.61(0.12)	0.61(0.12)
2006	0.64(0.12)	0.63(0.12)	0.63(0.12)	0.63(0.11)	0.63(0.11)	0.63(0.11)	0.63(0.11)	0.64(0.11)	0.64(0.11)
2007	0.67(0.10)	0.66(0.10)	0.65(0.10)	0.65(0.10)	0.65(0.10)	0.65(0.10)	0.65(0.10)	0.65(0.10)	0.66(0.10)
2008	0.65(0.08)	0.63(0.08)	0.61(0.08)	0.61(0.08)	0.60(0.08)	0.60(0.08)	0.60(0.09)	0.60(0.09)	0.60(0.09)
2009	0.66(0.08)	0.63(0.08)	0.61(0.07)	0.60(0.07)	0.59(0.07)	0.58(0.07)	0.58(0.08)	0.57(0.08)	0.57(0.08)
2010	0.65(0.07)	0.62(0.07)	0.59(0.07)	0.58(0.07)	0.57(0.07)	0.57(0.07)	0.57(0.07)	0.56(0.07)	0.56(0.08)
2011	0.61(0.08)	0.57(0.08)	0.54(0.07)	0.52(0.07)	0.51(0.07)	0.51(0.07)	0.50(0.07)	0.50(0.08)	0.49(0.08)
Panel B: Power-law distribution with an exponential cutoff (γ)									
2003	−0.87(0.07)	−0.81(0.07)	−0.79(0.07)	−0.77(0.08)	−0.76(0.08)	−0.75(0.09)	−0.74(0.09)	−0.74(0.09)	−0.74(0.10)
2004	−0.90(0.09)	−0.85(0.08)	−0.83(0.08)	−0.82(0.09)	−0.81(0.09)	−0.80(0.09)	−0.80(0.10)	−0.80(0.10)	−0.79(0.10)
2005	−0.80(0.11)	−0.77(0.11)	−0.76(0.11)	−0.75(0.11)	−0.74(0.12)	−0.74(0.12)	−0.74(0.12)	−0.74(0.12)	−0.73(0.12)
2006	−0.80(0.12)	−0.78(0.12)	−0.77(0.12)	−0.77(0.11)	−0.76(0.11)	−0.76(0.11)	−0.76(0.11)	−0.75(0.11)	−0.75(0.11)
2007	−0.83(0.10)	−0.81(0.10)	−0.79(0.10)	−0.78(0.10)	−0.78(0.10)	−0.77(0.10)	−0.77(0.10)	−0.76(0.10)	−0.76(0.10)
2008	−0.82(0.08)	−0.79(0.08)	−0.77(0.08)	−0.75(0.08)	−0.75(0.08)	−0.74(0.08)	−0.73(0.09)	−0.73(0.09)	−0.73(0.09)
2009	−0.84(0.08)	−0.79(0.08)	−0.77(0.07)	−0.75(0.07)	−0.74(0.07)	−0.73(0.07)	−0.72(0.08)	−0.71(0.08)	−0.71(0.08)
2010	−0.83(0.07)	−0.78(0.07)	−0.76(0.07)	−0.74(0.07)	−0.73(0.07)	−0.72(0.07)	−0.71(0.07)	−0.71(0.07)	−0.71(0.08)
2011	−0.79(0.08)	−0.74(0.08)	−0.71(0.07)	−0.69(0.07)	−0.67(0.07)	−0.67(0.07)	−0.66(0.07)	−0.65(0.08)	−0.65(0.08)
Panel C: q-exponential distribution (q)									
2003	1.20(0.07)	1.22(0.07)	1.22(0.07)	1.23(0.08)	1.23(0.08)	1.22(0.09)	1.22(0.09)	1.21(0.09)	1.21(0.10)
2004	1.17(0.09)	1.18(0.08)	1.18(0.08)	1.18(0.09)	1.18(0.09)	1.17(0.09)	1.16(0.10)	1.16(0.10)	1.15(0.10)
2005	1.23(0.11)	1.23(0.11)	1.23(0.11)	1.23(0.11)	1.23(0.12)	1.22(0.12)	1.22(0.12)	1.22(0.12)	1.21(0.12)
2006	1.23(0.12)	1.23(0.12)	1.22(0.12)	1.21(0.11)	1.21(0.11)	1.21(0.11)	1.20(0.11)	1.20(0.11)	1.19(0.11)
2007	1.21(0.10)	1.21(0.10)	1.21(0.10)	1.20(0.10)	1.20(0.10)	1.19(0.10)	1.19(0.10)	1.18(0.10)	1.18(0.10)
2008	1.23(0.08)	1.23(0.08)	1.24(0.08)	1.23(0.08)	1.23(0.08)	1.23(0.08)	1.23(0.09)	1.23(0.09)	1.22(0.09)
2009	1.22(0.08)	1.23(0.08)	1.24(0.07)	1.24(0.07)	1.24(0.07)	1.24(0.07)	1.24(0.08)	1.24(0.08)	1.24(0.08)
2010	1.23(0.07)	1.25(0.07)	1.25(0.07)	1.26(0.07)	1.26(0.07)	1.26(0.07)	1.26(0.07)	1.26(0.07)	1.26(0.08)
2011	1.26(0.08)	1.28(0.08)	1.29(0.07)	1.30(0.07)	1.31(0.07)	1.31(0.07)	1.31(0.07)	1.31(0.08)	1.31(0.08)
Panel D: Weibull distribution (ζ)									
2003	0.88(0.07)	0.85(0.07)	0.84(0.07)	0.83(0.08)	0.82(0.08)	0.82(0.09)	0.82(0.09)	0.81(0.09)	0.81(0.10)
2004	0.90(0.09)	0.88(0.08)	0.87(0.08)	0.86(0.09)	0.86(0.09)	0.86(0.09)	0.86(0.10)	0.86(0.10)	0.86(0.10)
2005	0.84(0.11)	0.82(0.11)	0.82(0.11)	0.81(0.11)	0.81(0.12)	0.81(0.12)	0.81(0.12)	0.81(0.12)	0.81(0.12)
2006	0.84(0.12)	0.83(0.12)	0.83(0.12)	0.83(0.11)	0.82(0.11)	0.82(0.11)	0.82(0.11)	0.82(0.11)	0.82(0.11)
2007	0.86(0.10)	0.85(0.10)	0.84(0.10)	0.84(0.10)	0.84(0.10)	0.83(0.10)	0.83(0.10)	0.83(0.10)	0.83(0.10)
2008	0.85(0.08)	0.83(0.08)	0.82(0.08)	0.82(0.08)	0.81(0.08)	0.81(0.08)	0.81(0.09)	0.81(0.09)	0.81(0.09)
2009	0.86(0.08)	0.84(0.08)	0.82(0.07)	0.81(0.07)	0.81(0.07)	0.80(0.07)	0.80(0.08)	0.79(0.08)	0.79(0.08)
2010	0.85(0.07)	0.83(0.07)	0.81(0.07)	0.80(0.07)	0.80(0.07)	0.79(0.07)	0.79(0.07)	0.79(0.07)	0.79(0.08)
2011	0.83(0.08)	0.79(0.08)	0.77(0.07)	0.76(0.07)	0.75(0.07)	0.75(0.07)	0.74(0.07)	0.74(0.08)	0.74(0.08)

$W(\Delta t|t)$ in equation (9) can be also applied to estimate the risk probability of coming extreme events in climate, seismic activities, stock returns, floods and so on.

Based on equation (10), we can estimate the risk probability of an incoming extreme event in the following period of Δt. If one sets $\Delta t = 1$, we will immediately have the information of hazard probability in next unit of time (minute here). Using a decision-making algorithm, we are allowed to predict whether there is a large volatility in the next minute (Bogachev and Bunde 2011). Specifically, we need to set a threshold Q_p for the hazard probability to trigger early warnings that a large volatility is about to occur. When the hazard probability exceeds Q_p, an alarm that a large volatility will occur during the next time point is activated. By comparing with actual extreme events, we can estimate the correct prediction and false alarm rates to evaluate the predicting performance. Instead of choosing a specific value of the hazard threshold, Q_p is varied in the range of [0, 1]. Such a strategy could avoid the situation of considering only one hazard threshold value: that a large value will increase the number of missing events and a small value will increase the number of false alarms. The plots of correct prediction rates with respect to false alarm rates from all possible hazard thresholds Q_p will give rise to the famous 'receiver operator characteristic' (ROC) curve (Bogachev and Bunde 2009a, 2009b, 2011, Bogachev *et al.* 2009), which graphically displays the forecasting efficiency.

To estimate the correct prediction and false alarm rates, we generate for a given Q_p two forecasting signals—alarms and non-alarms—at each time point. By comparing the forecasting signals with the real data, we obtain one of the four outcomes at each time point (Bogachev and Bunde 2011), (1) a correct prediction of a large volatility event, (2) a correct prediction of a non-large volatility event, (3) a missed event and (4) a false alarm. By recording in our testing records how many times each outcome occurs, we can estimate the correct prediction rate D and the false alarm rate A using

$$D = \frac{n_{11}}{n_{01} + n_{11}}, \quad A = \frac{n_{10}}{n_{00} + n_{10}}, \tag{12}$$

where n_{11} is the number of large volatility events that are correctly predicted, n_{00} the number of non-large volatility events that are correctly predicted, n_{01} the number of missed events and n_{10} the number of false alarms. All possible pairs of (D, A) will be obtained if we vary the Q_p range from 0 to 1.

By definition, the ROC curve will be $D = A = 1$ if $Q_p = 0$ and $D = A = 0$ if $Q_p = 1$. The ROC curve joins the point (0, 0) in the left bottom corner to the point (1, 1) in the right top corner. For the random guess outcome, $D = A$, a straight line is observed between the two corners. This occurs when there is no memory in the data. The area under the ROC curve is a measurement of predicting performance. In practical application, people are more interested in the predicting model with less false alarms. Hence, we define the following performance statistics to evaluate the predicting power,

$$\mathrm{AUC}_m = \int_0^{0.3} D(A)\mathrm{d}A, \tag{13}$$

where AUC_m is nothing but the area below the ROC curve in the range of $0 \leq A \leq 0.3$. Focusing A in this range is that the predictions are commonly useless if the false alarms are large. From this definition, the statistic AUC_m locates in the range of [0.045, 0.3]. The bottom limit corresponds to the random guess and the up limit is associated with the perfect prediction.

We estimate the volatility series $v(t)$ for each stock and regard the last year as an out-of-sample predicting period. The data not in the last year are considered as the in-sample training set, which is used to calibrate the model parameters. The forecasting of large volatilities is implemented as follows:

- The stretched exponential distribution parameter μ is estimated through fitting the recurrence intervals between large volatilities above a threshold Q associated with $\tau_Q = 100$ for the data in the in-sample training set.
- The events of large volatilities are identified in the out-of-sample predicting period based on the threshold Q in the in-sample calibrating period.
- The hazard probability $W(\Delta t|t)$ in the out-of-sample predicting period is determined based on the in-sample distribution parameter μ and the out-of-sample large volatilities.
- The correct prediction rates D and the false alarm rates A are calculated by comparing the predicting large volatilities with the actual out-of-sample large volatilities. The predicting large volatilities are obtained through making comparisons of the hazard probability $W(\Delta t|t)$ and the hazard threshold Q_p, which is varying from 0 to 1.
- The ROC curve is obtained through plotting D with respect A. The performance statistic AUC_m is estimated based on equation (13).

Figure 6(a) plots a subseries of the volatility in the out-of-sample predicting period and highlights events above threshold Q in the top panel. The risk probability $W(1|t)$ is shown in the bottom panel. Note that $W(1|t)$ decreases as time t elapsed from the last large volatility event increases. Threshold Q_p is plotted as a horizontal line to show the activating alarm process. By varying Q_p within a [0, 1] range, we obtain all pairs of (A, D). Figure 6(b) shows the ROC curves for two stocks. The two curves are above the dashed line $D = A$, indicating that our prediction is not random. The two curves do not overlap, indicating that the accuracy of this prediction algorithm varies for different stocks. We can observe that the AUC_m of stock 000001 is greater than that of stock 900956, which means that this model offers a relative better forecasting performance for stock 000001. We also calculate AUC_m for all stocks. Figure 6(c) shows the frequency plots of AUC_m. Note that the peak is centred at ≈ 0.08, about two times of a random prediction. We also find there are 22 stocks with $\mathrm{AUC}_m > 0.15$. The top three values are 0.2130, 0.2088 and 0.1993 for stocks 000557, 000529 and 000628, indicating that our forecasting algorithm can accurately predict the large volatilities of these three stocks. Previous research has indicated that the efficiency of the algorithm is primarily influenced by the linear and non-linear memory behaviour in the original volatilities (Bogachev and Bunde 2011). Our results could be explained by that the behaviour of stocks with stronger memory behaviours, such as volatility clustering and multifactality, could be more accurately predicted using our algorithm. Our algorithm only takes into consideration the probability distribution of recurrence intervals; if the memory behaviour of recurrence intervals is also included, we believe that its predictive accuracy would be greatly enhanced.

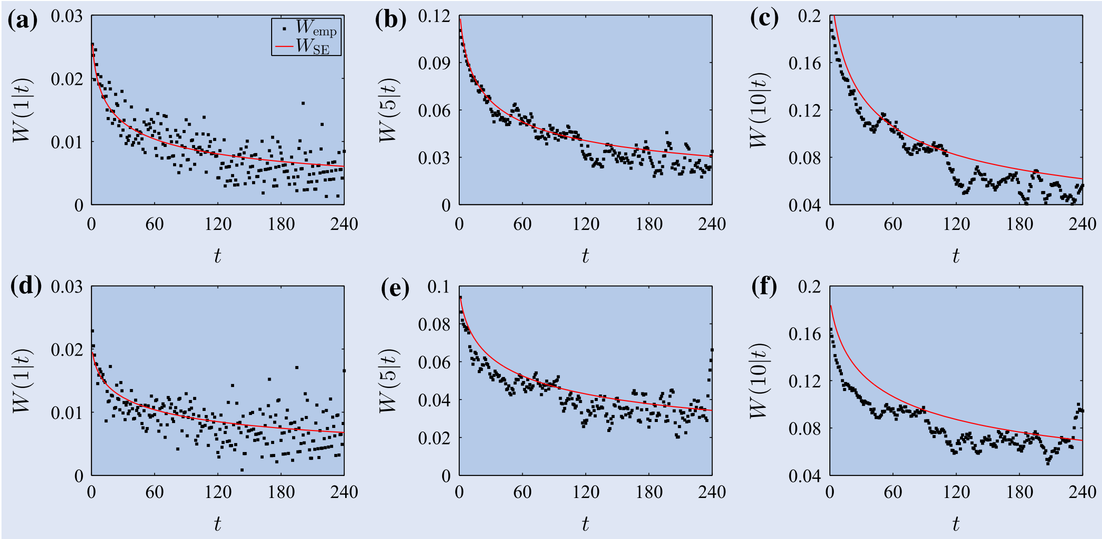

Figure 5. Comparison of the empirical hazard probability W_{emp} and the analytical hazard probability W_{SE} based on stretched exponential distribution for two stocks and $\Delta t = 1$, 5, and 10. (a)–(c) Stock 000001. (d)–(f) Stock 900956.

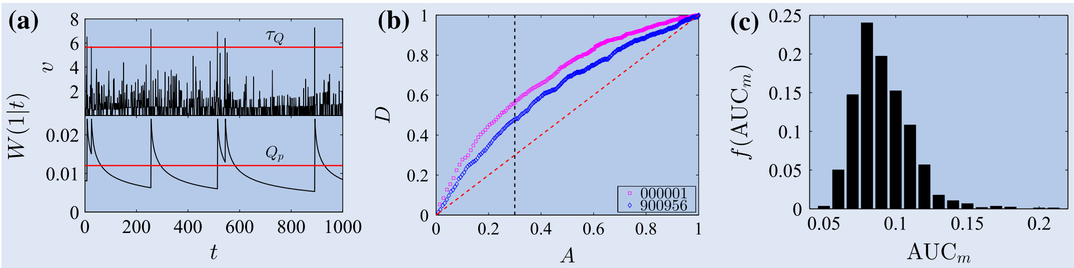

Figure 6. Prediction of large volatilities. (a) Plots of a representative volatility series in the top panels and hazard probability $W(1|\Delta t)$ in the bottom panels. (b) Plots of ROC curves for stock 000001 and stock 900956. (c) Distribution plots of AUC_m.

5. Conclusion

In this work, we have utilized a decision-making algorithm to forecast the occurrence of large volatilities in the Chinese stock markets based on the hazard probability, which is derived from the distribution of recurrence intervals between the volatilities exceeding a threshold Q. By fitting the volatility recurrence intervals by means of four candidate distributions and assessing the fitting validity by means of likelihoods and KS statistics, we have found that the volatility recurrence intervals are well approximated by a stretched exponential distribution. All the four distribution parameters (μ, γ, q and ζ) are found to be dependent of the mean recurrence time τ_Q, which has a one-to-one correspondence with the threshold Q, for all the stocks in our sample. Using a moving window analysis, we have found that the parameters are influenced by market status and exhibit a monotonic trend with the evolution of time, that μ and ζ decrease and γ and q increase. These behaviours are consistent with the inefficiency of the Chinese stock markets from 2007 to 2011.

Using the stretched exponential distribution formula, we have derived an analytical solution of the hazard probability $W(\Delta t|t)$ of the next large volatility event above the threshold Q within a short time interval Δt after an elapsed time t from the last large volatility above Q. This analytical solution $W(\Delta t|t)$ is in good agreement with the empirical risk probability estimated from real stock data. Based on a decision-marking algorithm, we have used the hazard probability to forecast large volatilities in the out-of-sample predicting period. We have found that the average statistic AUC_m, which corresponds to the area below the ROC curve in the range of $0 \le A \le 0.3$, is 0.091 for all stocks. We have also found that there are two stocks with $\mathrm{AUC}_m > 0.2$, indicating that our predicting algorithm is accurate in forecasting the large volatilities. Our findings may shed new light on our understanding of extreme volatility behaviour and may have potential applications in managing stock market risk.

Acknowledgements

We are grateful for Yu-Lei Wan for preprocessing the data.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

Z.-Q.J. and W.-X.Z. was supported by the National Natural Science Foundation of China [71131007 and 71532009], Shanghai 'Chen Guang' Project [2012CG34], Program for Changjiang Scholars and Innovative Research Team in University [IRT1028], China Scholarship Council [201406745014] and the Fundamental Research Funds for the Central Universities. A.C. acknowledges the support from Brazilian agencies FAPEAL [PPP20110902-011-0025-0069/60030-733/2011] and CNPq [PDE 20736012014-6]. H.E.S. was supported by NSF [Grants CMMI 1125290, PHY 1505000, and CHE-1213217] and by DOE Contract [DE-AC07-05Id14517].

ORCID

Wei-Xing Zhou http://orcid.org/0000-0002-8952-8228

References

Altmann, E.G. and Kantz, H., Recurrence time analysis, long-term correlations, and extreme events. *Phys. Rev. E*, 2005, **71**, 056106.

Bogachev, M.I. and Bunde, A., Memory effects in the statistics of interoccurrence times between large returns in financial record. *Phys. Rev. E*, 2008, **78**, 036114.

Bogachev, M.I. and Bunde, A., Improved risk estimation in multifractal records: Application to the value at risk in finance. *Phys. Rev. E*, 2009a, **80**, 026131.

Bogachev, M.I. and Bunde, A., On the occurrence and predictability of overloads in telecommunication networks. *EPL (Europhys. Lett.)*, 2009b, **86**, 66002.

Bogachev, M.I. and Bunde, A., On the predictability of extreme events in records with linear and nonlinear long-range memory: Efficiency and noise robustness. *Physica A*, 2011, **390**, 2240–2250.

Bogachev, M.I. and Bunde, A., Universality in the precipitation and river runoff. *EPL (Europhys. Lett.)*, 2012, **97**, 48011.

Bogachev, M.I., Eichner, J.F. and Bunde, A., Effect of nonlinear correlations on the statistics of return intervals in multifractal data sets. *Phys. Rev. Lett.*, 2007, **99**, 240601.

Bogachev, M.I., Eichner, J.F. and Bunde, A., On the occurence of extreme events in long-term correlated and multifractal data sets. *Pure Appl. Geophys.*, 2008a, **165**, 1195–1207.

Bogachev, M.I., Eichner, J.F. and Bunde, A., The effects of multifractality on the statistics of return intervals. *Eur. Phys. J. Spec. Top.*, 2008b, **161**, 181–193.

Bogachev, M.I., Kireenkov, I.S., Nifontov, E.M. and Bunde, A., Statistics of return intervals between long heartbeat intervals and their usability for online prediction of disorders. *New J. Phys.*, 2009, **11**, 063036.

Bollen, B. and Inder, B., Estimating daily volatility in financial markets utilizing intraday data. *J. Emp. Finance*, 2002, **9**, 551–562.

Bunde, A., Eichner, J.F., Havlin, S. and Kantelhardt, J.W., Return intervals of rare events in records with long-term persistence. *Physica A*, 2004, **342**, 308–314.

Bunde, A., Eichner, J.F., Kantelhardt, J.W. and Havlin, S., Long-term memory: A natural mechanism for the clustering of extreme events and anomalous residual times in climate records. *Phys. Rev. Lett.*, 2005, **94**, 048701.

Cai, S.M., Fu, Z.Q., Zhou, T., Gu, J. and Zhou P.L., Scaling and memory in recurrence intervals of internet traffic. *EPL (Europhys. Lett.)*, 2009, **87**, 68001.

Chicheportiche, R. and Chakraborti, A., A model-free characterization of recurrences in stationary time series. 2013. arXiv:1302.3704. Available online at: http://arxiv.org/abs/1302.3704

Chicheportiche, R. and Chakraborti, A., Copulas and time series with long-ranged dependencies. *Phys. Rev. E*, 2014, **89**, 042117.

Clauset, A., Shalizi, C.R. and Newman, M.E.J., Power-law distributions in empirical data. *SIAM Rev.*, 2009, **51**, 661–703.

Corral, A., Local distributions and rate fluctuations in a unified scaling law for earthquakes. *Phys. Rev. E*, 2003, **68**, 035102.

Dai, Y.H., Xie, W.J., Jiang, Z.Q., Jiang, G.J., and Zhou, W.X., Correlation structure and principal components in global crude oil market. *Emp. Econ.* Forthcoming.

Eichner, J.F., Kantelhardt, J.W., Bunde, A. and Havlin, S., Statistics of return intervals in long-term correlated records. *Phys. Rev. E*, 2007, **75**, 011128.

Greco, A., Sorriso-Valvo, L., Carbone, V. and Cidone, S., Waiting time distributions of the volatility in the Italian MIB30 index: Clustering or Poisson functions? *Physica A*, 2008, **387**, 4272–4284.

Guo, H.F., Wang, T.N., Li, Y.J. and Fung, H.G., Challenges to China's new stock market for small and medium-size enterprises: Trading price falls below the IPO price. *Technol. Econ. Dev. Econ.*, 2013, **19**, S409–S424.

Hallerberg, S., Altmann, E.G., Holstein, D. and Kantz, H., Precursors of extreme increments. *Phys. Rev. E*, 2007, **75**, 016706.

Hallerberg, S. and Kantz, H., Influence of the event magnitude on the predictability of an extreme event. *Phys. Rev. E*, 2008, **77**, 011108.

He, L.Y. and Chen, S.P., A new approach to quantify power-law cross-correlation and its application to crude oil markets. *Physica A*, 2011, **390**, 3806–3814.

Hussein, M.N. and Zhou, Z.G., The initial return and its conditional return volatility: Evidence from the Chinese IPO Market. *Rev. Pac. Basin Financ. Mark. Policy*, 2014, **17**, 1450022.

Jeon, W., Moon, H.T., Oh, G., Yang, J.S. and Jung, W.S., Return intervals analysis of the Korean stock market. *J. Korean Phys. Soc.*, 2010, **56**, 922–925.

Jiang, Z.Q., Chen, W. and Zhou, W.X., Scaling in the distribution of intertrade durations of Chinese stocks. *Physica A*, 2008, **387**, 5818–5825.

Jiang, Z.Q., Xie, W.J., Li, M.X., Podobnik, B., Zhou, W.X. and Stanley, H.E., Calling patterns in human communication dynamics. *Proc. Natl. Acad. Sci. USA*, 2013, **110**, 1600–1605.

Jiang, Z.Q., Xie, W.J. and Zhou, W.X., Testing the weak-form efficiency of the WTI crude oil futures market. *Physica A*, 2014, **405**, 235–244.

Jiang, Z.Q., Zhou, W.X., Sornette, D., Woodard, R., Bastiaensen, K. and Cauwels, P., Bubble diagnosis and prediction of the 2005–2007 and 2008–2009 Chinese stock market bubbles. *J. Econ. Behav. Org.*, 2010, **74**, 149–162.

Jung, W.S., Wang, F.Z., Havlin, S., Kaizoji, T., Moon, H.T. and Stanley, H.E., Volatility return intervals analysis of the Japanese market. *Eur. Phys. J. B*, 2008, **62**, 113–119.

Kaizoji, T. and Kaizoji, M., Power law for the calm-time interval of price changes. *Physica A*, 2004, **336**, 563–570.

Lee, J.W., Lee, K.E. and Rikvold, P.A., Waiting-time distribution for Korean stock-market index KOSPI. *J. Korean Phys. Soc.*, 2006, **48**, S123–S126.

Li, W., Wang, F.Z., Havlin, S. and Stanley, H.E., Financial factor influence on scaling and memory of trading volume in stock market. *Phys. Rev. E*, 2011, **84**, 046112.

Liu, C., Jiang, Z.Q., Ren, F. and Zhou, W.X., Scaling and memory in the return intervals of energy dissipation rate in three-dimensional fully developed turbulence. *Phys. Rev. E*, 2009, **80**, 046304.

Ludescher, J. and Bunde, A., Universal behavior of the interoccurrence times between losses in financial markets: Independence of the time resolution. *Phys. Rev. E*, 2014, **90**, 062809.

Ludescher, J., Tsallis, C. and Bunde, A., Universal behaviour of interoccurrence times between losses in financial markets: An analytical description. *EPL (Europhys. Lett.)*, 2011, **95**, 68002.

Meng, H., Ren, F., Gu, G.F., Xiong, X., Zhang, Y.J., Zhou, W.X., and Zhang, W, Effects of long memory in the order submission process on the properties of recurrence intervals of large price fluctuations. *EPL (Europhys. Lett.)*, 2012, **98**, 38003.

Meng, H., Xie, W.J., Jiang, Z.Q., Podobnik, B., Zhou, W.X. and Stanley, H.E., Systemic risk and spatiotemporal dynamics of the US housing market. *Sci. Rep.*, 2014, **4**, 3566.

Olla, P., Return times for stochastic processes with power-law scaling. *Phys. Rev. E*, 2007, **76**, 011122.

Pennetta, C., Distribution of return intervals of extreme events. *Eur. Phys. J. B*, 2006, **50**, 95–98.

Podobnik, B., Horvatic, D., Petersen, A.M. and Stanley, H.E., Cross-correlations between volume change and price change. *Proc. Natl. Acad. Sci. USA*, 2009, **106**, 22079–22084.

Politi, M. and Scalas, E., Fitting the empirical distribution of intertrade durations. *Physica A*, 2008, **387**, 2025–2034.

Qiu, T., Guo, L. and Chen, G., Scaling and memory effect in volatility return interval of the Chinese stock market. *Physica A*, 2008, **387**, 6812–6818.

Ren, F., Gu, G.F. and Zhou, W.X., Scaling and memory in the return intervals of realized volatility. *Physica A*, 2009a, **388**, 4787–4796.

Ren, F., Guo, L. and Zhou, W.X., Statistical properties of volatility return intervals of Chinese stocks. *Physica A* 2009b, **388**, 881–890.

Ren, F. and Zhou, W.X., Multiscaling behavior in the volatility return intervals of Chinese indices. *EPL (Europhys. Lett.)*, 2008, **84**, 68001.

Ren, F. and Zhou, W.X., Recurrence interval analysis of high-frequency financial returns and its application to risk estimation. *New J. Phys.*, 2010a, **12**, 075030.

Ren, F. and Zhou, W.X., Recurrence interval analysis of trading volumes. *Phys. Rev. E*, 2010b, **81**, 066107.

Saichev, A. and Sornette, D., "Universal" distribution of interearthquake times explained. *Phys. Rev. Lett.*, 2006, **97**, 078501.

Santhanam, M.S. and Kantz, H., Return interval distribution of extreme events and long-term memory. *Phys. Rev. E*, 2008, **78**, 051113.

Scalas, E., Gorenflo, R., Luckock, H., Mainardi, F., Mantelli, M. and Raberto, M., Anomalous waiting times in high-frequency financial data. *Quant. Finance*, 2004, **4**, 695–702.

Song, D.M., Tumminello, M., Zhou, W.X. and Mantegna, R.N., Evolution of worldwide stock markets, correlation structure, and correlation based graphs. *Phys. Rev. E*, 2011, **84**, 026108.

Sornette, D. and Knopoff, L., The paradox of the expected time until the next earthquake. *Bull. Seism. Soc. Am.*, 1997, **87**, 789–798.

Suo, Y.Y., Wang, D.H. and Li, S.P., Risk estimation of CSI 300 index spot and futures in China from a new perspective. *Econ. Model.*, 2015, **49**, 344–353.

Wang, F. and Wang, J., Statistical analysis and forecasting of return interval for SSE and model by lattice percolation system and neural network. *Comput. Ind. Eng.*, 2012, **62**, 198–205.

Wang, F.Z., Weber, P., Yamasaki, K., Havlin, S. and Stanley, H.E., Statistical regularities in the return intervals of volatility. *Eur. Phys. J. B*, 2007, **55**, 123–133.

Wang, F.Z., Yamasaki, K., Havlin, S. and Stanley, H.E., Scaling and memory of intraday volatility return intervals in stock markets. *Phys. Rev. E*, 2006, **73**, 026117.

Wang, F.Z., Yamasaki, K., Havlin, S. and Stanley, H.E., Indication of multiscaling in the volatility return intervals of stock markets. *Phys. Rev. E*, 2008, **77**, 016109.

Wang, F.Z., Yamasaki, K., Havlin, S. and Stanley, H.E., Multifactor analysis of multiscaling in volatility return intervals. *Phys. Rev. E*, 2009, **79**, 016103.

Xie, W.J., Jiang, Z.Q. and Zhou, W.X., Extreme value statistics and recurrence intervals of NYMEX energy futures volatility. *Econ. Model.*, 2014, **36**, 8–17.

Yamasaki, K., Muchnik, L., Havlin, S., Bunde, A. and Stanley, H.E., Scaling and memory in volatility return intervals in financial markets. *Proc. Natl. Acad. Sci. USA*, 2005, **102**, 9424–9428.

Yamasaki, K., Muchnik, L., Havlin, S., Bunde, A. and Stanley, H.E., Scaling and memory in return loss intervals: Application to risk estimation, in *Proceedings of the Practical Fruits of Econophysics*, edited by H. Takayasu, pp. 43–51, 2006 (Springer-Verlag: Berlin).