World Scientific
www.worldscientific.com

Testing of binary regime switching models using squeeze duration analysis

Milan Kumar Das[*] and Anindya Goswami[†]

Mathematical Sciences
Indian Institute of Science Education and Research
Pune, India

Abstract

We have developed a statistical technique to test the model assumption of binary regime switching extension of the geometric Brownian motion (GBM) model by proposing a new discriminating statistics. Given a time series data, we have identified an admissible class of the regime switching candidate models for the statistical inference. By performing several systematic experiments, we have successfully shown that the sampling distribution of the test statistics differs drastically, if the model assumption changes from GBM to Markov modulated GBM, or to semi-Markov modulated GBM. Furthermore, we have implemented this statistics for testing the regime switching hypothesis with Indian sectoral indices.

Keywords: Empirical volatility; regime switching GBM; time series analysis; parameter inference.

1. Introduction

In the work of Black, Scholes, and Merton in 1973, a geometric Brownian motion (GBM) model was used to explain the dynamics of risky asset prices. This was subsequently reexamined and extended by several authors to fill the gaps between the empirical observations and the consequences from theoretical models. In this context, stochastic models following the work of Heston (1993) appear promising. However, the calibration of such models and their extensions became increasingly challenging. Simultaneously the regime switching models following the works of

---
[†]Corresponding author.

Email addresses: [*]das.milan2@gmail.com, [†]anindya@iiserpune.ac.in

Hamilton (1989) and Di Masi *et al.* (1994), also earned popularity in the literature and research in this direction continue to grow. Regime switching GBM is one such extension. Various authors assumed regime switching models and addressed theoretical option pricing and portfolio optimization problems following Di Masi *et al.* (1995). For details, the readers may refer to Basak *et al.* (2011), Bo *et al.* (2010), Fan *et al.* (2014), Jobert and Rogers (2006), Li and Ma (2013), Lian *et al.* (2016), Papin and Turinici (2014), Siu *et al.* (2008), Su *et al.* (2012), Swishchuk *et al.* (2014), Wang *et al.* (2016) and Yuen and Yang (2009). This list is merely indicative and not exhaustive by any means. However, the statistical verification of effectiveness of the regime switching models for an asset price data has not yet been satisfactorily addressed in the literature. Nevertheless, using a particular regime switching model, some stylized facts in the daily return had been explained in Jan and Bulla (2006) and Rydén *et al.* (1998). It is important to note that, in a regime switching dynamics, each of the key market parameters evolves as a finite state pure jump process which might or might not be a Markov chain. Replacement of a Markov chain by a semi-Markov chain results in an immediate extension. As per our knowledge, there is no comprehensive statistical analysis which helps discriminating among the cases of GBM, Markov-modulated GBM (MMGBM) and semi-Markov modulated GBM (SMGBM) for modeling a given asset price time series data. However, in Bulla and Bulla (2006), the authors have studied the stylized facts of returns of daily time series and have shown that the hidden semi-Markov models describe the stylized facts including the slow decay in the autocorrelation function (ACF) of the squared returns in a better way than what MMGBM model does.

The main difference between a Markov chain and a semi-Markov chain is in its instantaneous transition rate. It is just a constant matrix for Markov case whereas for the semi-Markov chain it is a matrix valued function on the positive real numbers. Despite the presence of many computational and theoretical studies on the SMGBM model for asset prices, the use of SMGBM is not as popular as it deserves to be. In recent works (Ghosh and Goswami, 2009; Goswami *et al.*, 2016; Goswami and Nandan, 2016; Das *et al.*, 2018), it is shown that the option price equation in SMGBM model can be solved with similar ease to the MMGBM counterpart. Furthermore, the portfolio optimization problem in SMGBM leads to the same optimal strategy as in the MMGBM case (Das *et al.*, 2018). Hence the SMGBM model is as appealing as the special case of MMGBM in terms of the mathematical tractability. So far the applicability is concerned, the SMGBM models are superior to the Markovian counterpart for its greater flexibility in fitting the inter transition times. Since such flexibility in model fitting directly leads to the improvements in derivative pricing, the statistical comparison between the MMGBM and SMGBM samples becomes particularly important. For the

comparison purpose, we propose a discriminating statistics whose sampling distribution varies drastically, under the regime switching assumption, with varying choices of instantaneous rate parameter. Our discriminating statistics is constructed using some descriptive statistics of squeeze duration of Bollinger band, which seems to be the most natural approach. The sampling distribution of the descriptive statistics of occupation measure of Bollinger band under a particular model hypothesis need not have a nice form and thus one may not be able to identify that as one of a few known distributions. Hence, the inference cannot be done analytically.

In spite of the lack of mathematical tractability, one may surely obtain empirical distribution of the statistics using a reliable simulation procedure. This is a standard approach in such circumstances. We refer Theiler and Prichard (1996) for further details. In Theiler *et al.* (1992) this is termed as the typical realization surrogate data approach. The surrogate data approach is generally perceived one of the most powerful method for testing of hypothesis. The available algorithms to generate surrogate can be categorized into two classes namely typical realizations and constrained realization (Schreiber and Schmitz, 2000). This paper considers the typical realization approach.

In this article, we study the geometric Brownian motion, or GBM in short, and its generalizations by using binary regime switching. By a binary regime, we mean that there exists an unobserved two-state stochastic process whose movement causes in changing the market parameters, i.e., drift and volatility coefficients. Let S_t denote the asset price at time t, under the binary regime switching GBM. Then $\{S_t\}_{t\geq 0}$ satisfies $S_0 > 0$ and

$$dS_t = S_t(\mu(X_t)dt + \sigma(X_t)dB_t), \tag{1.1}$$

where $\{X_t\}_{t\geq 0}$ is a $\{1, 2\}$-valued stochastic process and $\mu(X_t), \sigma(X_t)$ are the drift and the volatility coefficients and B_t is the standard Brownian motion. In general $\{X_t\}_{t\geq 0}$ is chosen to be Markov or semi-Markov process. The GBM model is a special case of this where $\mu(1) = \mu(2)$ and $\sigma(1) = \sigma(2)$. Since the GBM model is quite popular among derivative market practitioners, we have compared the distribution of the discriminating statistics for a surrogate data produced from geometric Brownian motion (GBM) with that for an empirical data. We also have compared the sampling distributions of statistics for the surrogate data produced from Markov modulated GBM and semi-Markov modulated GBM, respectively. We have considered eighteen sectoral indices from Indian stock market for this purpose. Corresponding to each time series data, we have first figured out the relevant composite null hypotheses and then the sampling distribution of the test statistics under each of those. Next a systematic comparison of the sampling distributions of the proposed test statistics is carried out for each of the composite

null hypotheses based on GBM, MMGBM, SMGBM models. As an outcome, the GBM hypothesis is rejected for each time series data with nearly 100% confidence. The MMGBM hypotheses are also rejected with more than 95% confidence for some time series. However, the SMGBM hypothesis could not be rejected at such confidence level.

The rest of this paper is arranged in the following manner. We dedicate the second section in describing the statistics based on the squeeze of the Bollinger bands. In this section we also recall some definitions from statistics and introduce the notion of squeeze durations in our context. The empirical study has been done in Sec. 3. Sections 4 and 5 contain concluding remark and acknowledgments, respectively.

2. Discriminating Statistics Based on Squeeze Duration

In this section, a discriminating statistics is proposed whose sampling distribution varies drastically, under the regime switching assumption, with varying values of instantaneous rate parameter. The discriminating statistics is taken as vector valued where every component is a descriptive statistics of squeeze duration of Bollinger band. This section is dedicated in describing the statistics and the numerical methods for obtaining its sampling distributions. The actual numerical experiments are deferred to the next section. This section is organized in four subsections.

2.1. Bollinger band

Keltner channel and Bollinger bands based on the empirical volatility are the most popular indicators for trading. John Bollinger introduced the concept of Bollinger band for pattern recognition in 1980s. Bollinger bands provide a time varying interval for any financial time series data. The end points of the intervals are computed based on the moving average and the moving sample standard deviation of the past data of fixed window size. Now we present a formal definition of the Bollinger bands of an asset.

Definition 2.1. A Bollinger band of a given time series data consists of three lines on the time series plot, computed based on immediate lag values of fixed length n say. The middle line is the moving average of the time series with window size n. The upper and the lower lines are exactly $k\sigma$ unit away from the middle line where k is a fixed constant and σ is the sample standard deviation obtained from the last n numbers of lag values.

It is important to note that the main focus of the Bollinger band is to capture the fluctuation, to be more precise, the volatility coefficient of the time series. Hence,

the closeness of the upper and lower line is termed as squeeze, and is the indication of low volatility of a particular time series. On the contrary, when the boundaries of band are far from each other, that corresponds to a high volatility. For more details about Bollinger bands and squeeze we refer to Bollinger (2001).

2.2. *p*-Squeeze durations

In this paper, we consider the Bollinger bands of the simple return of a financial time series. We introduce some important notations and definitions which would be used subsequently. Let $S = \{S_k\}_{k=1}^N$ denote an equispaced financial time series. The simple return of S is defined by

$$r_k := \frac{S_k - S_{k-1}}{S_{k-1}}, \quad k = 2, 3, \dots, N. \tag{2.1}$$

Definition 2.2 $(\hat{\mu}, \hat{\sigma})$. By fixing the window size as n, the moving average $\{m_k\}_{k=n+1}^N$ and the sample standard deviation $\{\sigma_k\}_{k=n+1}^N$ are given by

$$m_k := \frac{1}{n} \sum_{i=0}^{n-1} r_{k-i}, \tag{2.2}$$

$$\sigma_k := \sqrt{\frac{1}{n-1} \sum_{i=0}^{n-1} (r_{k-i})^2 - \frac{n}{n-1} m_k^2}, \tag{2.3}$$

for $k \geq n + 1$. The empirical volatility $\hat{\sigma} = \{\hat{\sigma}_k\}_{k=n+1}^N$ is given by $\hat{\sigma}_k := \frac{\sigma_k}{\sqrt{\Delta}}$ for all $k \geq n + 1$, where Δ is the length of the time step in year unit. Similarly, the empirical drift $\hat{\mu} = \{\hat{\mu}_k\}_{k=n+1}^N$ is given by $\hat{\mu}_k := \frac{m_k}{\Delta}$ for all $k \geq n + 1$.

Definition 2.3. Let $y = \{y_k\}_{k=1}^m$ be a random sample of a real valued random variable. Then the empirical cumulative distribution function or ecdf $\hat{F}_y$ is defined as

$$\hat{F}_y(x) := \frac{1}{m} \sum_{k=1}^m \mathbb{1}_{[0, \infty)}(x - y_k),$$

where given a subset A, $\mathbb{1}_A$ denotes the indicator function of A.

Definition 2.4 (*p*-percentile). Let $\hat{F}_y$ be the ecdf of $y = \{y_k\}_{k=1}^m$. Then for any $p \in (0, 1)$, the *p*-percentile of y, denoted by $\hat{F}_y^{\leftarrow}(p)$, is defined as

$$\hat{F}_y^{\leftarrow}(p) := \inf \{x | \hat{F}_y(x) \geq p\}.$$

For mathematical tractability we use a particular percentile of $\hat{\sigma}$ as threshold in defining the squeeze of the Bollinger band.

Definition 2.5 (*p*-squeeze). Given a $p \in (0, 1)$, an asset is said to be in *p*-squeeze at kth time step if the empirical volatility $\hat{\sigma}_k$, as defined above, is not more than $\hat{F}^{\leftarrow}_{\hat{\sigma}}(p)$.

We introduce the sojourn times of the *p*-squeeze below.

Definition 2.6. For a fixed $p \in (0, 1)$ and a given time series $\{S_k\}^N_{k=1}$, let $\{(a_i, b_i)\}^\infty_{i=1}$ be an extended real valued double sequence given by

$$\begin{cases} a_0 = 0 \\ b_{i-1} := \min\{k \geq a_{i-1} | \hat{\sigma}_k > \hat{F}^{\leftarrow}_{\hat{\sigma}}(p)\} \\ a_i := \min\{k \geq b_{i-1} | \hat{\sigma}_k \leq \hat{F}^{\leftarrow}_{\hat{\sigma}}(p)\}, \end{cases}$$

for $i = 1, 2, \ldots$ and by following the convention of $\min \varnothing = +\infty$, where $\hat{\sigma}$ is as in Definition 2.2. Then the sojourn time durations for the *p*-squeezes are $\{d_i\}^L_{i=1}$, where $d_i := b_i - a_i$ and $L := \max\{i | b_i < \infty\}$, provided $L \geq 1$.

We note that one must multiply each d_i by Δ to obtain the squeeze durations in year unit. We would consider the finite sequence $\{d_i\}^L_{i=1}$ as a single object. In particular, we call d_i as the ith entry of the *p*-squeeze duration or *p*-SqD in short for the given time series $\{S_k\}^N_{k=1}$. We call L to be the length of *p*-SqD.

Remark 2.7. In a reasonably large and practically relevant time series data, the length of *p*-SqD is considerably small. Hence a non-parametric estimation of the entries of *p*-SqD using empirical cdf is not practicable as that would have a high standard error. Hence, only a collection of some descriptive statistics such as mean($\bar{d}$), standard deviation(s), skewness(ν), kurtosis(κ) of *p*-SqD can reliably be obtained and compared.

2.3. A discriminating statistics

We first consider a discrete time version of continuous time theoretical asset price model with the time step identical to that of the time series data. We note that, for that theoretical model, the corresponding *p*-SqD is a random sequence with random length. However, the corresponding descriptive statistics as above would constitute a random vector of fixed length whose sampling distribution would be sought for. A comparison of $(\bar{d}, s, \nu, \kappa)$ with respect to that sampling distribution would be the central idea for statistical inference. However, this should not lead to the only criterion for rejecting a model. Of course there are many other natural criteria for the same. Those criteria are typically considered as constraints on

parameterization of the class of models of our interest. Here we illustrate with an example. If we restrict ourselves to all possible MMGBM models with two regime as in Bulla and Bulla (2006), then

$$\Theta = \{(\mu(1), \sigma(1), \lambda_1, \mu(2), \sigma(2), \lambda_2) | \mu(i) \in \mathbb{R}, \sigma(i) > 0, \lambda(i) > 0, i = 1, 2\} \tag{2.4}$$

is the class of parameters, where $\mu(i)$ and $\sigma(i)$ are the drift and the volatility coefficients respectively and $\begin{pmatrix} -\lambda_1 & \lambda_1 \\ \lambda_2 & -\lambda_2 \end{pmatrix}$ is the rate matrix for the Markov chain. Therefore the estimation problem boils down to a constrained minimization on a set $\mathscr{C} \subset \Theta$ of the following functional $f : \Theta \to \mathbb{R}$ given by

$$f(\theta) = E^\theta \big[(\bar{d}^\theta - \bar{d})^2 + (s^\theta - s)^2 + (\nu^\theta - \nu)^2 + (\kappa^\theta - \kappa)^2 \big], \tag{2.5}$$

where $(\bar{d}^\theta, s^\theta, \nu^\theta, \kappa^\theta)$ is the descriptive statistics vector of a member with parameter $\theta \in \Theta$, provided the minimizer exists.

The main difficulty in taking $\mathscr{C} = \Theta$ is its time complexity due to a large scope of parameter values. We introduce a fixed set of constraints, p-admissible class ($\mathscr{C}_p$-class), which is a subclass of all possible regime switching models.

Definition 2.8 ($\mathscr{C}_p$**-class**).   Given a time series data and a fixed $p \in (0, 1)$, a regime switching model is said to be in $\mathscr{C}_p$-class of models if the model satisfies the following properties.

(i) The long run average of drift coefficient matches with the time average of empirical drift $\hat{\mu}$.
(ii) The long run average of volatility parameters matches with the time average of empirical volatility $\hat{\sigma}$.
(iii) The long run proportion of time that the volatility process stays below $\hat{F}_{\hat{\sigma}}^{\leftarrow}(p)$ is p, provided the volatility process is not constant.

In view of Remark 2.7, we construct a discriminating statistics $\mathbf{T} = (T_1, T_2, \ldots, T_r)$ using r number of descriptive statistics of the p-SqD. To be more specific we choose $T_1 := \frac{1}{L} \sum_{i=1}^{L} d_i$, $T_2 := \sqrt{\frac{1}{L-1} \sum_{i=1}^{L} (d_i - T_1)^2}$, $T_3 := \frac{\frac{1}{L} \sum_{i=1}^{L} (d_i - T_1)^3}{T_2^3}$, $T_4 := \frac{\frac{1}{L} \sum_{i=1}^{L} (d_i - T_1)^4}{T_2^4}$, etc. Although our test statistics is based on squeeze durations which are amenable to capture the sojourn times of regime transitions but it is not at all obvious that it would indeed be successful to capture those unobserved switchings. The main difficulty lies in the fact that a larger moving window size (n) in defining $\hat{\sigma}$ ignores more number of intermittent transitions and a smaller window size corresponds to higher standard error. So far window size is concerned there is a popular choice of window size by

practitioners, i.e., $n = 20$ for computing the empirical volatility. In view of these, we fix $n = 20$ now onward in the definition of $\mathbf{T}$.

Next we describe the procedure, adopted in this paper, of obtaining the sampling distribution of $\mathbf{T}$ under binary regime switching model hypothesis.

2.4. Sampling distribution of the statistics

In this section we give a detailed description of numerical computation of sampling distribution of $\mathbf{T}$ statistics under the null hypothesis using Monte Carlo method, which is popularly known as typical surrogate approach following Theiler *et al.* (1992). It is important to note that the hypothesis testing, relevant to us, is of composite type (Yuen and Yang, 2009). The main purpose is to test a meaningful composite null hypothesis. The procedure is as follows

(a) Given a time series S, the p-admissible class $\mathscr{C}_p$ under the null hypothesis is identified. A non-empty subclass $\mathscr{A}$ of $\mathscr{C}_p$ is fixed.
(b) For each $\theta \in \mathscr{A}$, B number of time series $\{X^1, X^2, \ldots, X^B\}$ are simulated from the corresponding model θ with the same time step as in S. We call these, the surrogate data of S corresponding to θ.
(c) Let $\mathbf{t}^* := \mathbf{T}(S)$ be the value of $\mathbf{T}$ of the observed data S and $\mathbf{t} :=$ $\{\mathbf{t}^1, \mathbf{t}^2, \ldots, \mathbf{t}^B\}$ be the values of $\mathbf{T}$ for the surrogate data $\{X^1, X^2, \ldots, X^B\}$, where $\mathbf{t}^i = (t_1^i, t_2^i, \ldots, t_r^i) = \mathbf{T}(X^i)$.
(d) By keeping a two sided test in mind we define α_r^θ in the following manner:

$$\alpha_r^\theta := 2 \min_{j \le r} g_B \left(\sum_{i=1}^B \mathbb{1}_{[0,\infty)}(t_j^* - t_j^i) \right),$$

where $g_B(x) := \frac{x \wedge (B-x)}{B}$, and $\mathbf{t}^* = (t_1^*, t_2^*, \ldots, t_r^*)$.
(e) Therefore the α-value for the test of the class $\mathscr{A}$ is given by

$$\alpha_r = \max_{\theta \in \mathscr{A}} \alpha_r^\theta.$$

(f) We reject the hypothesis that S is a sample from a model in the class with confidence $100(1 - \alpha_r)\%$, provided α_r is reasonably small.

Remark 2.9. It is important to note, the above method has a pathetic limitation due to the "curse of dimensionality." Or in other words for a given model θ, the probability of observing the value of α_r^θ to be smaller than a very small value is not so small when r is large. But it is well known that the curse is not so fatal for the dimension r less than five. Therefore we restrict ourselves in four-dimensional testing.
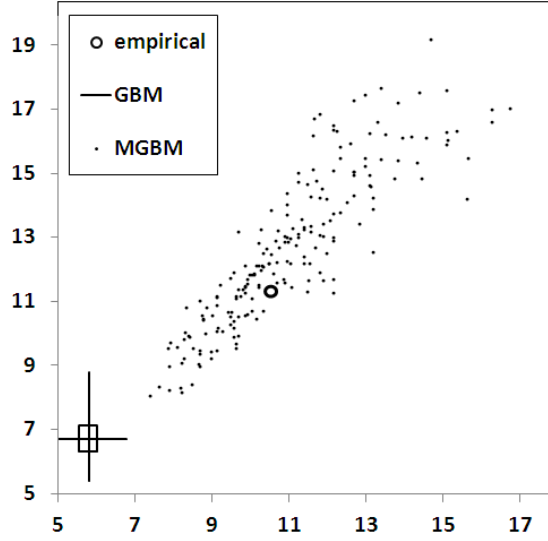
Fig. 1. Plot of (T_1, T_2) for observation and various different surrogate data.

For the purpose of illustration here we consider a specific time series S and $r = 2$, i.e., the test statistics $\mathbf{T} = (T_1, T_2)$. In the following plot Fig. 1, the values T_1 and T_2 are plotted against the horizontal and vertical axes, respectively. The position of the point $t^* = \mathbf{T}(S)$ is denoted as a circle in the plot. Under the null hypothesis of GBM, the $\mathscr{C}_p$-class turns out to be singleton (see Sec. 3.1 for details). The sampling distribution of $\mathbf{T}$ under the null hypothesis is computed by setting $B = 200$ and that is presented using a two-dimensional box plot. Furthermore, under the null hypothesis of MMGBM, followed by fixing $B = 1$, and a subclass $\mathscr{A} := \{\frac{1}{\lambda_1} \in (\frac{1}{20}\mathbb{N}) \cap [5, 15], \sigma(1) = \hat{F}_{\hat{\sigma}}^{\leftarrow}(p)\}$ of $\mathscr{C}_p$, the values of $\mathbf{T}$ are plotted as dots.

In the following section, we implement the ideas developed here for more specific choice of models and discuss the implementation issues in details.

3. Empirical Study

For empirical study, we consider, 5-min data of several Indian stock indices from 1st December, 2016 to 30th June, 2017. Assuming there are 250 trading days in a year and 6 hours of trading in each day, we set $\Delta = \frac{5}{250 \times 360} \approx 5.5 \times 10^{-5}$. We fix $p = 15\%$ throughout this section. The components of t^* for each index data are given in the Table 1 below. Every row of the table corresponds to an index, which is mentioned in the second column with their id's in the first column. The third

Table 1. Values of discriminating statistics ($p = 15\%$) of a 5-min data of Indian indices during 1st December 2016 to 30th June 2017.

Data	Indices	Occurence	t_1^*	t_2^*	t_3^*	t_4^*
I01	NIFTY 100	159	10.52	11.31	1.17	3.41
I02	NIFTY 200	160	10.45	11.18	1.29	3.79
I03	NIFTY 50	155	10.78	11.00	1.08	3.28
I04	NIFTY 500	152	11.01	11.40	1.20	3.63
I05	NIFTY BANK	159	10.52	11.62	1.39	4.03
I06	NIFTY COMMODITY	169	9.89	10.49	1.47	4.59
I07	NIFTY ENERGY	168	9.96	11.44	1.59	4.80
I08	NIFTY FIN. SER.	168	9.95	10.72	1.46	4.39
I09	NIFTY FMCG	178	9.40	10.15	1.58	5.01
I10	NIFTY INFRA	174	9.61	11.70	1.72	5.41
I11	NIFTY IT	159	10.52	11.36	1.19	3.35
I12	NIFTY MEDIA	173	9.66	9.49	1.19	3.77
I13	NIFTY METAL	188	8.89	10.53	1.92	6.52
I14	NIFTY MNC	178	9.40	10.63	1.54	4.67
I15	NIFTY PHARMA	175	9.56	11.13	1.59	4.69
I16	NIFTY PSE	148	11.29	12.73	1.28	3.76
I17	NIFTY REALTY	177	9.45	10.49	1.83	6.02
I18	NIFTY SERVICE SEC.	172	9.72	11.13	1.33	3.77

column gives the value of L, the number of observations of p-squeeze duration of each index data.

3.1. GBM hypothesis

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space on which $\{W_t\}_{t\geq 0}$ be a Brownian motion. The stock price, modeled by a geometric Brownian motion (or GBM in short) is given by

$$dS_t = S_t(\mu dt + \sigma dW_t) \quad t \geq 0, \ S_0 > 0, \tag{3.1}$$

where μ and σ are the drift and the volatility coefficients respectively. Equation (3.1) has a strong solution of the form

$$S_t = S_0 \exp\left(\mu t - \frac{1}{2}\sigma^2 t + \sigma W_t\right), \quad t \geq 0. \tag{3.2}$$

3.1.1. *Surrogate data*

It is important to note that, the $\mathscr{C}_p$-class is singleton as μ and σ are, by using Definition 2.8 (i)–(ii), $\mu = \bar{\hat{\mu}}$ and $\sigma = \bar{\hat{\sigma}}$, where the bar sign represents the time average.

Let S_0 be the initial price of a stock consisting of N number of data points. Let $\{0 = t_1 < t_2 < \ldots < t_N\}$ be a partition of time interval of the observed data series, where $t_{i+1} - t_i = \Delta$ for $i = 1, 2, \ldots, N - 1$ and Δ be the length of time step in year unit. Then the $\mathscr{C}_p$-class of surrogate GBM can be generated by using the discretized version of (3.2) which is given by

$$S_{t_{i+1}} = S_{t_i} \exp\left(\left(\bar{\tilde{\mu}} - \frac{1}{2}\bar{\tilde{\sigma}}^2\right)\Delta + \bar{\tilde{\sigma}}Z_i\right), \quad S_{t_1} = S_0, \tag{3.3}$$

where $\{Z_i | i = 1, \ldots, N - 1\}$ are independent and identically distributed (i.i.d.) normal random variables with mean 0 and variance Δ. We further emphasize that the step size δ, used for the discretization is identical to the step size in the observed time series. We use this convention throughout this paper.

3.1.2. *Testing of hypothesis*

We intend to test whether the value of $\mathbf{T}$ of observed index prices are outliers of $\mathbf{T}$ values coming from GBM models. For each index in Table 1, we set our null hypothesis,

$$H_0 : \text{the time series is in } \mathscr{C}_p\text{-class of GBM.}$$

We again recall that the $\mathscr{C}_p$-class is indeed singleton. The following figures illustrate results from all 18 indices. Figure 2 plots T_1 and Fig. 3 plots T_2 only. Each box plot is obtained by simulating the GBM model from $\mathscr{C}_p$-class 200 times. The triangle plots are the representative for original data of all the indices. Here we see that the triangles appear far from the box plots.
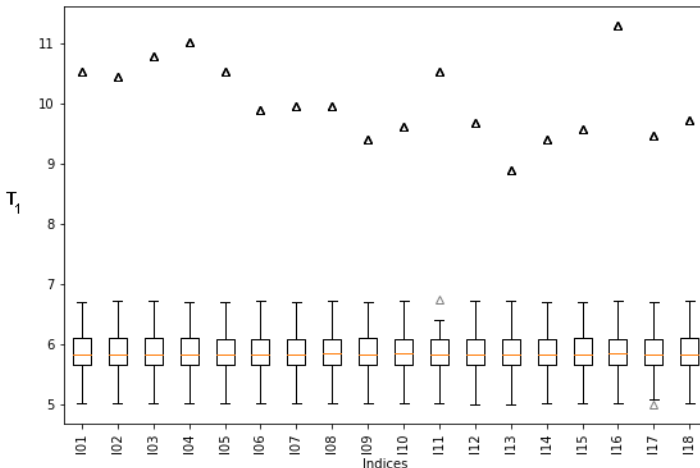

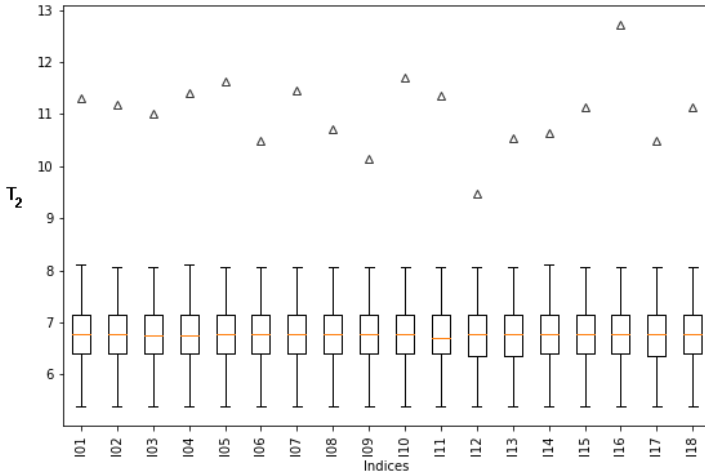Fig. 2. Sampling distribution of T_1 under GBM hypothesis.

Fig. 3. Sampling distribution of T_2 under GBM hypothesis.

Therefore, Figs. 2 and 3 indicate a strong rejection for the null hypothesis of GBM model. We continue our investigation with binary regime switching Markov modulated geometric Brownian motion in the following subsection.

3.2. Markov modulated GBM hypothesis

In this subsection, we study the testing of hypothesis of $\mathscr{C}_p$-class of MMGBM assumption with binary regimes. In this case, Eq. (1.1) is dependent on a two state Markov process $\{X_t\}_{t \geq 0}$. It is important to note that a continuous time Markov chain can be characterized by its instantaneous transition rate matrix.

3.2.1. *Surrogate data*

In this subsection, we restrict our investigation in a particular subclass $\mathscr{A}$ of $\mathscr{C}_p$ which are embedded in

$$\Theta = \{\theta = (\mu(1), \sigma(1), \lambda_1, \mu(2), \sigma(2), \lambda_2) | \mu(i) \in \mathbb{R}, \sigma(i) > 0, \lambda_i > 0, i = 1, 2\}. \tag{3.4}$$

Here the transition rate matrix for the Markov chain is given by

$$\Lambda := \begin{pmatrix} -\lambda_1 & \lambda_1 \\ \lambda_2 & -\lambda_2 \end{pmatrix}.$$

Therefore, the sojourn time distribution of state i is $\text{Exp}(\lambda_i)$ for $i = 1, 2$. Now since $\mathscr{A} \subset \mathscr{C}_p \subset \Theta$, by using Definition 2.8(iii), we have

$$\frac{\frac{1}{\lambda_1}}{\frac{1}{\lambda_1} + \frac{1}{\lambda_2}} = p. \tag{3.5}$$

Using Definition 2.8(i), (3.5) the drift coefficients $\mu(i)$ satisfy the following relation:

$$p\mu(1) + (1 - p)\mu(2) = \bar{\mu}. \tag{3.6}$$

Also using Definition 2.8(ii) and (3.5) the volatility coefficients $\sigma(i)$ have the relation below

$$p\sigma(1) + (1 - p)\sigma(2) = \bar{\sigma}. \tag{3.7}$$

After a simple computation (3.5) becomes

$$\lambda_1 = \left(\frac{1}{p} - 1\right)\lambda_2. \tag{3.8}$$

Thus, $\mathscr{C}_p \subset \Theta$ is the set of six parameters satisfying Eqs. (3.6)–(3.8). We choose $\mathscr{A}$ by fixing $\mu(1) = \mu(2)$ and $\sigma(1) = \hat{F}_{\bar{\sigma}}^{\leftarrow}(p)$. Thus, $\mathscr{A}$ is a subset of the solution space of five equations in six unknowns, or in other words, $\mathscr{A}$ can be viewed as a one-parameter family of models. Next to generate surrogate data, we need to discretize the MMGBM model corresponding to each member of $\mathscr{A}$. To this end we discretize (1.1) and perform Monte Carlo simulation. The discretization of MMGBM surrogate of the form (1.1) is given by

$$S_{t_{i+1}} = S_{t_i} \exp\left(\left(\mu(X_i) - \frac{1}{2}\sigma^2(X_i)\right)\Delta + \sigma(X_i)Z_i\right),$$
$$X_{i+1} = X_i - (-1)^{X_i}P_i, \tag{3.9}$$

where $\{P_i | i = 1, \ldots, N - 1\}$ are independent to Z_j for all j and for each given X_i, the conditional distribution of P_i is independent of $\{P_1, P_2, \ldots, P_{i-1}\}$ and follows *Bernoulli*$(\lambda_{X_i}\Delta)$, a Bernoulli random variable with $Prob(P_i = 1 | X_i) = \lambda_{X_i}\Delta$, provided $\Delta \ll \min\{1/\lambda_i | i = 1, 2\}$. Here for each i, Z_i is as in (3.3). For sufficiently small Δ, the above discrete model approximates the continuous time MMGBM model. A proof of this fact is excluded here.

3.2.2. *Testing of hypothesis*

We intend to test whether the values of **T** of observed index prices are outliers of **T** values coming from MMGBM models. For each index in Table 1, we set our

null hypothesis,

$$H_0 : \text{the time series is in the class } \mathscr{A} \text{ of MMGBM}.$$

To test H_0, we adopt the typical realization surrogate data approach and consider the discrete model (3.9). For every $\theta \in \mathscr{A}$, we perform Monte Carlo simulation two hundred times ($B = 200$). Then we record the α_1, α_2, α_3 and α_4 values for each index in Table 2. In order to enhance understanding of the composite hypothesis, here we consider the first time series I01 to illustrate the sampling distribution of $\mathbf{T}$ for some models in $\mathscr{A}$ using the following four 2D projection plots. In each of Figs. 4–7, the circle plot represents t^*, the $\mathbf{T}$ value of I01. Furthermore, there are three two-dimensional box plots corresponding to $\frac{1}{\lambda_1}$ equal to 5, 10 and 15, respectively. By abuse of notation we write $\frac{1}{\lambda_1}$ as θ. It seems from Fig. 4 that, the least square estimate of $\frac{1}{\lambda_1}$ should lie in between 5 and 10, so far (T_1, T_2) is concerned. Indeed as in Table 2, the α_2 value (0.395), which depends only on T_1 and T_2, is considerably large. However, Fig. 5 implies that the least square estimate of $\frac{1}{\lambda_1} > 15$ by considering (T_3, T_4) only. On the other hand, each of Figs. 6 and 7 implies that the circle plot is an outlier. Thus no combination of parameters in the composite hypothesis can explain the observed data. We can also

Table 2. The α-vales for all the indices.

|	Index	MMGBM				SMGBM			
		α_1	α_2	α_3	α_4	α_1	α_2	α_3	α_4
1	I01	0.490	0.395	0.045	0.040	0.495	0.485	0.225	0.095
2	I02	0.490	0.400	0.050	0.045	0.500	0.455	0.195	0.105
3	I03	0.470	0.420	0.050	0.035	0.495	0.470	0.230	0.100
4	I04	0.435	0.395	0.055	0.040	0.500	0.455	0.180	0.105
5	I05	0.415	0.395	0.055	0.040	0.495	0.415	0.195	0.105
6	I06	0.465	0.390	0.055	0.040	0.485	0.415	0.200	0.115
7	I07	0.430	0.430	0.060	0.040	0.500	0.420	0.205	0.100
8	I08	0.475	0.420	0.050	0.035	0.495	0.455	0.235	0.100
9	I09	0.455	0.420	0.085	0.050	0.495	0.425	0.220	0.105
10	I10	0.460	0.390	0.055	0.040	0.485	0.425	0.200	0.115
11	I11	0.455	0.420	0.055	0.040	0.490	0.430	0.190	0.100
12	I12	0.480	0.395	0.050	0.050	0.500	0.470	0.210	0.115
13	I13	0.490	0.405	0.050	0.040	0.500	0.470	0.215	0.105
14	I14	0.430	0.395	0.055	0.040	0.485	0.415	0.195	0.100
15	I15	0.490	0.410	0.050	0.035	0.480	0.480	0.235	0.100
16	I16	0.435	0.395	0.055	0.040	0.495	0.425	0.200	0.105
17	I17	0.470	0.410	0.045	0.030	0.495	0.400	0.195	0.095
18	I18	0.425	0.395	0.055	0.040	0.490	0.430	0.205	0.100
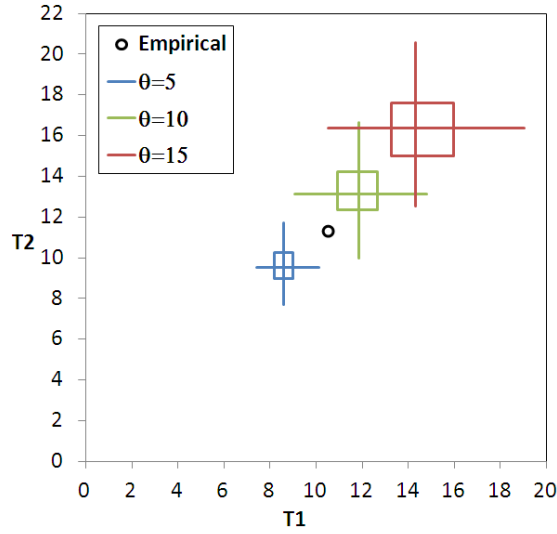
Fig. 4. T_1 and T_2 under MMGBM hypothesis.

observe another very important feature from these four plots. The sampling distributions of the statistics change monotonically with the change of parameter θ. Such monotonic sensitivity of the statistics is crucial for successful inference.

From Table 2 it is evident that H_0 can be rejected for each index with α value 5% or smaller. Finally, since all the α_4 are less than 5%, in the MMGBM column of Table 2, we reject the null hypothesis with 95% confidence for each index data.
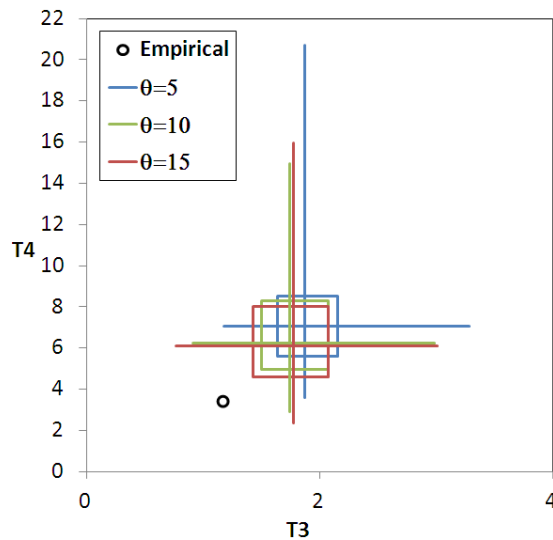
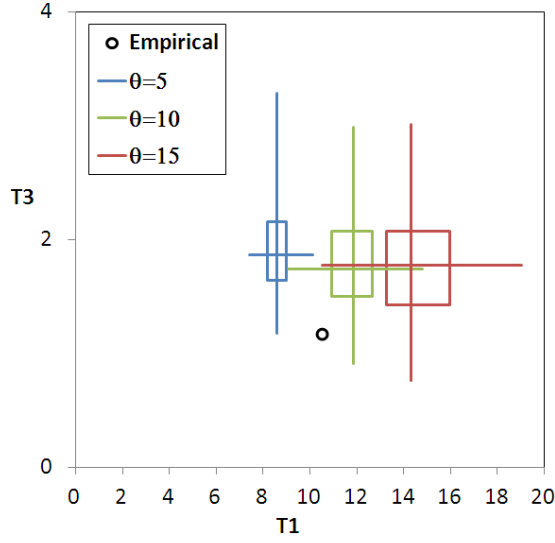Fig. 5. T_3 and T_4 under MMGBM hypothesis.

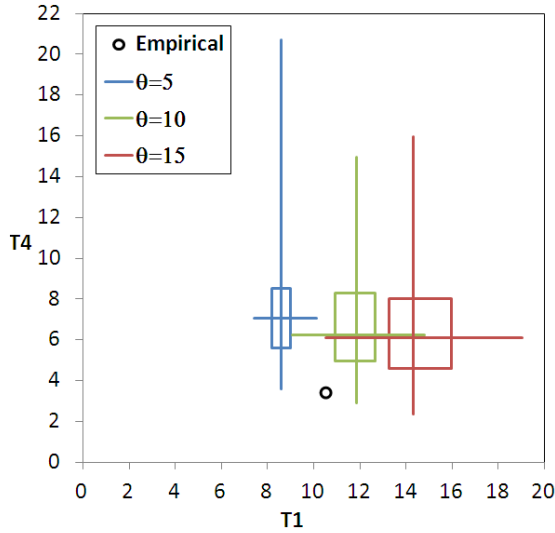Fig. 6. T_1 and T_3 under MMGBM hypothesis.

Fig. 7. T_1 and T_4 under MMGBM hypothesis.

3.3. Semi-Markov modulated GBM

In this subsection, we consider the SMGBM assumption with binary regimes. In this case, Eq. (1.1) is dependent on a two state semi-Markov process $\{X_t\}_{t \geq 0}$. A semi-Markov process can be characterized by its instantaneous transition rate function. Alternatively a semi-Markov process can also be specified by fixing the

transition probability matrix of the embedded chain and the conditional distribution of holding time at each state.

3.3.1. *Surrogate data*

A subclass of the following all class of models

$$\Theta = \{\theta = (\mu(1), \sigma(1), \lambda_1(\cdot), \mu(2), \sigma(2), \lambda_2(\cdot)) | \mu(i) \in \mathbb{R},$$
$$\sigma(i) > 0, \lambda_i(\cdot) > 0, i = 1, 2\}.$$

would be considered now. For every θ in $\mathscr{C}_p \subset \Theta$, $\mu(i)$ and $\sigma(i)$ satisfy Eqs. (3.6) and (3.7). The transition rate matrix for the semi-Markov chain is a matrix valued function on $[0, \infty)$, given by

$$\Lambda(y) := \begin{pmatrix} -\lambda_1(y) & \lambda_1(y) \\ \lambda_2(y) & -\lambda_2(y) \end{pmatrix} \quad \forall y \in [0, \infty).$$

Now for illustration purpose, $\mathscr{A}$ is chosen in the following manner. The holding time distribution of the state i is $\Gamma(k_i, \lambda_i)$ for $i = 1, 2$, where $\Gamma(k_i, \lambda_i)$ denote the gamma distribution with shape k_i and rate λ_i. Then it follows from Ghosh and Saha (2011) that $\lambda_i(y)$ is the hazard rate of $\Gamma(k_i, \lambda_i)$ and is given by $\lambda_i(y) = \frac{\lambda_i^{k_i} y^{k_i-1} e^{-\lambda_i y}}{\Gamma(k_i) - \gamma(k_i, \lambda_i y)}$, where γ is the lower incomplete gamma function. Since the expectation of $\Gamma(k_i, \lambda_i)$ is $\frac{k_i}{\lambda_i}$, it follows from Definition 2.8(iii), that

$$\frac{\frac{k_1}{\lambda_1}}{\frac{k_1}{\lambda_1} + \frac{k_2}{\lambda_2}} = p,$$

i.e.,

$$\frac{k_2}{\lambda_2} = \left(\frac{1}{p} - 1\right)\frac{k_1}{\lambda_1}.$$

In addition to these, as before, we further assume that $\mu(1) = \mu(2)$, $\sigma(1) = \hat{F}_{\hat{\sigma}}^{\leftarrow}(p)$ and $k_1 = k_2$. Thus $\mathscr{A}$ is the solution space of six equations in eight unknowns or in other words $\mathscr{A}$ is a two parameter subfamily of Θ. For drawing samples from each member of $\mathscr{A}$ using Monte Carlo simulation, we first discretize (1.1). The discretization scheme for SMGBM surrogate of (1.1) is given by

$$S_{t_{i+1}} = S_{t_i} \exp\left(\left(\mu(X_i) - \frac{1}{2}\sigma^2(X_i)\right)\Delta + \sigma(X_i) Z_i\right),$$
$$X_{i+1} = X_i + (-1)^{X_i} P_i, \tag{3.10}$$
$$Y_{i+1} = (Y_i + i\Delta)(1 - P_i),$$

where $\{Z_i\}_i$ are as in (3.3) and $\{P_i | i = 1, \dots, N - 1\}$ are independent to Z_j for all j and for each given pair (X_i, Y_i), the conditional distribution of P_i is independent of $\{P_1, P_2, \dots, P_{i-1}\}$ and follows *Bernoulli*$(\lambda_{X_i}(Y_i)\Delta)$, a Bernoulli random variable with $Prob(P_i = 1 | X_i, Y_i) = \lambda_{X_i}(Y_i)\Delta$, provided $\Delta \ll \min\{1/\lambda_i(y) | y \geq 0, i = 1, 2\}$. This discretization is obtained from the semi-martingale representation of the semi-Markov process, as in Ghosh and Saha (2011). The readers are referred to Ghosh and Saha (2011) for more details about this representation of semi-Markov process.

3.3.2. *Testing of hypothesis*

We set our null hypothesis for all index

$$H_0 : \text{the time series is in the class } \mathscr{A} \text{ of SMGBM.}$$

From Table 2 below, H_0 cannot be rejected for any index with a significant level of confidence. Hence, we cannot reject the superset also. Or in other words, we cannot reject the composite hypothesis that the data, under study, is drawn from a SMGBM population.

4. Conclusion

In this paper, we have developed a statistical technique to test the hypothesis of the binary regime switching extensions of GBM. The proposed statistical procedure involves four steps. In the first step we discretize the continuous time model. Monte Carlo simulations of the discrete model is carried out with specific combination of parameters in the second step. In the third step we compute the sampling distribution of the statistics under the model hypothesis for various combinations of parameters. In the final step we compare the statistics value of the observed data with the sampling distributions obtained in the third step. Despite mathematical intractability, we present the procedure systematically with sufficient rigor. In view of the unavailability of any reported investigation on inference of regime switching models, we have taken the most natural and simplistic approach to the problem. We have successfully observed that even with this simple approach, the sampling distributions of the statistics exhibit a monotonic sensitivity to the model parameters. Thus, the proposed statistics is suitable for the inference problem. The paper is written for a very broad class of readers by eliminating several technical jargons. This work, in principle, gives a framework and opens up opportunity of many research investigation concerning inference of regime switching dynamics arising in several different applied fields.

Acknowledgments

References

Basak, GK, MK Ghosh and A Goswami (2011). Risk minimizing option pricing for a class of exotic options in a Markov-modulated market, *Stochastic Analysis and Applications*, 29(2), 259–281.

Bo, L, Y Wang and X Yang (2010). Markov-modulated jump-diffusions for currency option pricing, *Insurance: Mathematics and Economics*, 46(3), 461–469.

Bollinger, J (2001). Bollinger on Bollinger Bands, *McGraw Hill Professional*.

Bulla, J and I Bulla (2006). Stylized facts of financial time series and hidden semi-Markov models, *Computational Statistics & Data Analysis*, 51(4), 2192–2209.

Das, MK, A Goswami and TS Patankar (2018). Pricing derivatives in a regime switching market with time inhomogeneous volatility, *Stochastic Analysis Applications*, 36(4), 700–725.

Das, MK, A Goswami and N Rana (2018). Risk sensitive portfolio optimization in a jump diffusion model with regimes, *SIAM Journal of Control Optimisation*, 56, 1550–1576.

Di Masi, GB, YM Kabanov and WJ Runggaldier (1995). Mean-variance hedging of options on stocks with Markov volatilities, *Theory of Probability & Its Applications*, 39(1), 172–182.

Fan, K, Y Shen, TK Siu and R Wang (2014). Pricing foreign equity options with regime-switching, *Economic Modelling*, 37, 296–305.

Ghosh, MK and A Goswami (2009). Risk minimizing option pricing in a semi-Markov modulated market, *SIAM Journal on Control and Optimization*, 48(3), 1519–1541.

Ghosh, MK and S Saha (2011). Stochastic processes with age-dependent transition rates, *Stochastic Analysis and Applications*, 29(3), 511–522.

Goswami, A, J Patel and P Shevgaonkar (2016). A system of non-local parabolic PDE and application to option pricing, *Stochastic Analysis and Applications*, 34(5), 893–905.

Goswami, A and S Nandan (2016). Convergence of estimated option price in a regime switching market, *Indian Journal of Pure and Applied Mathematics*, 47(2), 169–182.

Hamilton, JD (1989). A new approach to the economic analysis of nonstationary time series and the business cycle, *Econometrica: Journal of the Econometric Society*, 357–384.

Heston, SL (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options, *Stochastic Analysis Applications*, 6(2), 327–343.

Jobert, A and LCG Rogers (2006). Option pricing with Markov-modulated dynamics, *SIAM Journal on Control and Optimization*, 44(6), 2063–2078.

Li, J and S Ma (2013). Pricing options with credit risk in Markovian regime-switching markets, *Journal of Applied Mathematics*, 2013, 621371.

Lian, Y-M, J-H. Chen and S-L Liao (2016). Option pricing on foreign exchange in a Markov-modulated, incomplete-market economy, *Finance Research Letters*, 16, 208–219.

Papin, T and G Turinici (2014). Prepayment option of a perpetual corporate loan: The impact of the funding costs, *International Journal of Theoretical and Applied Finance*, 17(4), 1450028.

Rydén, T, T Terasvirta and S Åsbrink (1998). Stylized facts of daily return series and the hidden Markov model, *Journal of Applied Econometrics*, 217–244.

Schreiber, T and A Schmitz (2000). Surrogate time series, *Physica D: Nonlinear Phenomena*, 142(3), 346–382.

Siu, TK, H Yang and JW Lau (2008). Pricing currency options under two-factor Markov-modulated stochastic volatility models, *Insurance: Mathematics and Economics*, 43(3), 295–302.

Su, X, W Wang and K-S Hwang (2012). Risk-minimizing option pricing under a Markov-modulated jump-diffusion model with stochastic volatility, *Statistics & Probability Letters*, 82(10), 1777–1785.

Swishchuk, A, M Tertychnyi and R Elliott (2014). Pricing currency derivatives with Markov-modulated Lévy dynamics, *Insurance: Mathematics and Economics*, 57, 67–76.

Theiler, J, S Eubank, A Longtin, B Galdrikian and JD Farmer (1992). Testing for nonlinearity in time series: The method of surrogate data, *Physica D: Nonlinear Phenomena*, 58(1–4), 77–94.

Theiler, J and D Prichard (1996). Constrained-realization Monte-Carlo method for hypothesis testing, *Physica D: Nonlinear Phenomena*, 94(4), 221–235.

Wang, W, Z Jin, L Qian and X Su (2016). Local risk minimization for vulnerable European contingent claims on nontradable assets under regime switching models, *Stochastic Analysis and Applications*, 34(4), 662–678.

Yuen, FL and H Yang (2009). Option pricing in a jump-diffusion model with regime switching, *ASTIN Bulletin: The Journal of the IAA*, 39(2), 515–539.