A simulation method for finite non-stationary time series

[a]Yuzhi Cai* , [b]Jie Huang, [b]Yu Tang and [b]Guixia Zhou
[a]*Swansea University, UK*
[b]*Soochow University, China*

Abstract

In this paper we propose a novel simulation method which enables us to obtain a large number of simulated time series cheaply. The developed method can be applied to any non-stationary time series of finite length and it guarantees that not only the marginal distributions but also the autocorrelation structures of observed and simulated time series are the same. Extensive simulation studies have been conducted to check the performance of our method and to assess if the overall dynamics of the observed time series is preserved by the simulated realizations. The developed simulation method has also been applied to the real size data of cocoon filament, which can be reeled from a cocoon produced by a silkworm. Very good results have been achieved in all the cases considered in the paper.

Key words: Autocorrelation structure, cocoon filament, marginal distribution, simulation, non-stationary time series.

1 Introduction

Simulation of time series plays an important role in many areas such as coastal engineering and raw reeling silk industry. In the statistical literature on time series, the focus has often been given to the ability to generate correctly short Gaussian time series from a given vector ARMA (VARMA) process. Several methods may be used to obtain a simulated time series in practice. The simplest method is to fit a VARMA model to a time series, then to generate a simulated time series from the fitted model. To do this, a Monte Carlo method or a Quasi Monte Carlo method could be used (see, for example, Li and Winker, 2003). Barone (1987) described a method for generating independent realizations of a VARMA process which involves recasting the ARMA model in a state space form and allows for an exact generation of the initial values of the simulation algorithm. Shea (1988) discussed a direct

*Address for correspondence: Dr Yuzhi Cai, College of Business, Economics and Law, Swansea University, Swansea, SA2 8PP, United Kingdom. Email: y.cai@swansea.ac.uk

method of computing the initial state covariance matrix required by the simulation method. However, these types of approaches usually suffer from the following problems, i.e. either the marginal distributions of the time series will differ from the ones requested, or the autocorrelation patterns will differ from the ones expected. In practice, we would like the simulated time series to be as similar to the observed time series as possible. Hence it would be of great interest and is very challenging to guarantee both the marginal distributions and the autocorrelation structures of the simulated and the observed time series to be the same.

Recently, Cai et al. (2008) and Cai (2010) generalized Cario and Nelson (1998), Deler and Nelson (2001) and Biller and Nelson's (2003, 2008) work to obtain a simulated multivariate time series that has the required autocorrelation structure and marginal distribution. However, their methods are suitable for stationary time series only.

In practice, we often have non-stationary time series. For example, in raw reeling silk industry, it is important to study the size of cocoon filament, which can be reeled from a cocoon produced by a silkworm. The size, measured by using the international standard unit dtex (mass in grams per 10000 meters), represents the thickness of the filament. It is well known that the size series of cocoon filament (SSCF) is a non-stationary stochastic series with finite length (see, for example, Fei and Bay, 2005, 2009). A large number of SSCF need to be studied by the researchers in raw silk reeling industry for manufacturing raw silk with the required standards. However, it is very expensive to collect many size series. Therefore, simulating size time series of good quality is of great importance, and it is also an even more challenging task because of the non-stationarity of the size series. Current approach to this challenging problem is to establish time varying parameter autoregressive time series models and then to use the fitted model to obtain simulated time series. See, for example, Fei and Bay (2005, 2009) and references therein. Although the simulated time series obtained from these methods are non-stationary with the required autocorrelation structure, there is no guarantee on the marginal distributions.

In this paper, we propose a simulation method for generating non-stationary time series with the required autocorrelation structure and the required marginal distributions, so that the quality of the simulated time series can be improved significantly.

The arrangements of the paper is as follows. The new simulation method is presented in Section 2. We carried out extensive simulation studies on the performance of our method, the results of which can be found in Section 3. An application to real size series of cocoon filament is given in Section 4. Further comments and conclusions are given in Section 5.

2 The simulation method

Let $y_1, \ldots, y_T$ be a non-stationary time series of length T. Let y_{it} $(t = 1, \ldots, T)$ be the ith $(i = 1, \ldots, I)$ realization of the time series in the time period $[1, T]$, which is independent of all other realizations of the time series. We need to obtain simulated time series with high quality based on the observed time series so that further analysis can be carried out.

To understand the above settings, let us consider the SSCF. In this case, the underlying time series corresponds to the size series of cocoon filament, i corresponds to the ith

silkworm. So we have I measured size series using I silkworms. Furthermore, under certain conditions silkworms are assumed to perform independently (see Fei and Bay (2005, 2009)). Note that as it is very expensive to obtain a very large number of SSCF, the value of I (i.e. the maximum number of SSCF that we can afford) is usually smaller than that the researchers would need for a required analysis. So simulated SSCF with high quality based on the data obtained from I silkworms can be very useful in such situations. This real problem also motivated the work of this paper.

Let $F_t(y)$ be the marginal distribution function of y_t at time t, $\rho_{t\,\tau} = corr(y_t,\, y_\tau)$ the correlation between y_t and y_τ with $\rho_{t\,t} = 1$ and $\rho_{t\,\tau} = \rho_{\tau\,t}$. As the time series is non-stationary, we have $F_t(y) \neq F_\tau(y)$ and $corr(y_t,\, y_{t+h}) \neq corr(y_\tau,\, y_{\tau+h})$ if $t \neq \tau$ and $h \neq 0$. The proposed simulation method consists of the following several steps.

Step 1. Estimate the marginal distributions and the sample autocorrelations.

Step 2. Define a base process z_t.

Step 3. Estimate the autocorrelation structure of the base process.

Step 4. Estimate the coefficients of the base process.

Step 5. Use the base process to obtain a simulated time series and then transform it to the required time series.

The details about each step of the simulation method are given below.

Step 1. Estimate the marginal distributions and the sample autocorrelations.

The marginal distribution function $F_t(y)$ ($t = 1, \ldots, T$) may be approximated by the empirical distribution or by fitting a proper probability model to the data y_{it} ($i = 1, \ldots, I$).

The autocorrelation $\rho_{t\,\tau}$ may be estimated by

$$\hat{\rho}_{t\,\tau} = \frac{1}{(I-1)\hat{s}_t\hat{s}_\tau}\sum_{i=1}^{I}(y_{i\,t} - \hat{\mu}_t)(y_{i\,\tau} - \hat{\mu}_\tau),$$

where

$$\hat{\mu}_t = \frac{1}{I}\sum_{i=1}^{I} y_{i\,t}\,, \quad \hat{s}_t^2 = \frac{1}{I-1}\sum_{i=1}^{I}(y_{i\,t} - \hat{\mu}_t)^2, \quad t = 1, \ldots, T.$$

Step 2. Define a base process z_t.

$$\begin{aligned} &z_1 = \epsilon_1, \\ &z_t = \theta_{1\,t-1}z_{t-1} + \theta_{2\,t-1}z_{t-2} + \cdots + \theta_{t-1\,t-1}z_1 + \epsilon_t, \quad t = 2, \ldots, T, \end{aligned} \tag{1}$$

where ϵ_t ($t = 1, \ldots, T$) are independently and normally distributed with

$$\epsilon_1 \sim N(0, 1), \qquad \epsilon_t \sim N(0, \sigma_t^2),$$

where

$$\sigma_t^2 = 1 - \theta_{1\,t-1}r_{t-1\,t} - \theta_{2\,t-1}r_{t-2\,t} - \cdots - \theta_{t-1\,t-1}r_{1\,t},$$

and $r_{t\,\tau}$ is the correlation between z_t and z_τ. Furthermore, ϵ_t is independent of z_τ for $\tau < t$. For the base process we have the following result.

Theorem 1 *The marginal distribution of the base process is standard normal, i.e.* $z_t \sim N(0,1)$ *(*$t = 1, \ldots, T$*).*

The proof can be found in the Appendix. The following result is a standard one and we present it as a Lemma.

Lemma 1 *Let* $Z \sim N(0,1)$*, and let* $Y = F^{-1}(\Phi(Z))$*, where* F^{-1} *is the inverse function of a properly defined distribution function, and* Φ *is the standard normal distribution function. Then* Y *is a random variable with a distribution function defined by* F*.*

Therefore, if we can simulate a time series z_t from the base process, then by letting $y_t = F_t^{-1}(\Phi(z_t))$ we can obtain a simulated time series y_t with the required marginal distributions. That means, we only need to guarantee that the simulated y_t process has the required autocorrelation structure, which can be achieved by estimating $r_{t\tau}$ in Step 3.

Step 3. Estimate the autocorrelation structure of the base process.

It is noticed that

$$\rho_{t\,\tau} = corr(y_t,\, y_\tau) = \frac{E(y_t y_\tau) - E(y_t)E(y_\tau)}{\sqrt{var(y_t)}\sqrt{var(y_\tau)}} = \frac{E(y_t y_\tau) - \mu_t \mu_\tau}{s_t s_\tau}\,. \tag{2}$$

where

$$\mu_t = E(y_t), \quad s_t = \sqrt{var(y_t)}, \quad t = 1, \ldots, T.$$

It is also noticed that

$$E(y_t y_\tau) = E\left(F_t^{-1}(\Phi(z_t))F_\tau^{-1}(\Phi(z_\tau))\right)$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F_t^{-1}(\Phi(z_t))F_\tau^{-1}(\Phi(z_\tau))\phi_{r_{t\,\tau}}(z_t,\, z_\tau)dz_t dz_\tau\,,$$

where $\phi_{r_{t\,\tau}}(z_t,\, z_\tau)$ is the bivariate normal density function of z_t and z_τ with mean 0 and correlation $r_{t\,\tau}$.

It is seen that $E(y_t y_\tau)$ is a function of $r_{t\tau}$ only, which appears in the function $\phi_{r_{t\,\tau}}(z_t,\, z_\tau)$. Thus, the problem of determining $r_{t\,\tau}$ that gives the desired autocorrelations for the y_t process reduces to independently solving the equations given by (2). For a stationary process, Cario and Nelson (1996) proved that $E(y_t y_\tau)$ is a nondecreasing function of $r_{t\tau}$, and under very mild conditions on the marginal distribution of y_t, $E(y_t y_\tau)$ is also continuous. For the non-stationary time series considered in this paper, we have similar results:

Theorem 2 $E(y_t y_\tau)$ *is nondecreasing for* $-1 \le r_{t\tau} \le 1$*. Furthermore, if there exists* $\varepsilon > 0$ *such that* $E(|y_t y_\tau|^{1+\varepsilon}) < \infty$ *for all values of* $-1 \le r_{t\tau} \le 1$*, then* $E(y_t y_\tau)$ *is continuous for* $-1 \le r_{t\tau} \le 1$*.*

See the Appendix for the proof. Theorem 2 guarantees that our numerical procedure developed below will converge.

By rearranging (2) we have

$$\rho_{t\,\tau} s_t s_\tau + \mu_t \mu_\tau = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F_t^{-1}(\Phi(z_t))F_\tau^{-1}(\Phi(z_\tau))\phi_{r_{t\,\tau}}(z_t,\, z_\tau)dz_t dz_\tau\,. \tag{3}$$

Replacing $\rho_{t\,\tau}$, s_t, s_τ, μ_t and μ_τ by the corresponding sample estimates, we can solve (3) for $r_{t\,\tau}$ numerically. We could use the method proposed by Cai et al. (2008) to estimate $r_{t\,\tau}$. However, as now we are dealing with non-stationary time series, the number of equations given by (3) can be very large if T is large. Therefore we propose the following new but much more efficient method for estimating $r_{t\,\tau}$.

Note that $-1 \leq r_{t\,\tau} \leq 1$. Let N be a large positive integer and $-1 < r_1 < \ldots < r_N < 1$, where r_j $(j = 1, \ldots, N)$ are equally spaced. Let v_k and w_k be independent samples from $N(0, 1)$, where $k = 1, \ldots, M$ and M is also a large positive integer (in this paper, we take $M = 5000$). Let $a_i = r_i/\sqrt{1 - r_i^2}$, $u_k^i = (a_i v_k + w_k)/\sqrt{a_i^2 + 1}$, where $i = 1, \ldots, N$. Finally, let

$$A_{t\,\tau}^i = \frac{1}{M} \sum_{k=1}^{M} F_t^{-1}(\Phi(v_k)) F_\tau^{-1}(\Phi(u_k^i)), \quad t = 2, \ldots, T,\ \tau = 1, \ldots, t-1.$$

Then

$$A_{t\,\tau}^i \approx \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F_t^{-1}(\Phi(z_t)) F_\tau^{-1}(\Phi(z_\tau)) \phi_{r_i}(z_t,\, z_\tau) dz_t,$$

and the approximation can be very good if M is large enough. Hence this step of the method includes the following several sub-steps.

- For fixed τ, simulate v_k and u_k as mentioned above. Calculate $A_{t\,\tau}^i$ $(t = \tau+1, \ldots, T)$. Note that the same values of $F_\tau^{-1}(\Phi(u_k^i))$ $(i = 1, \ldots, N)$ are required for all t. Therefore they only need to be calculated once when $t = \tau + 1$.
- Let $A_{t\,\tau}^{(1)} < A_{t\,\tau}^{(2)} < \cdots < A_{t\,\tau}^{(N)}$ be the ordered $A_{t\,\tau}^i$ $(t = \tau + 1, \ldots, T)$, and $r^{(j)}$ be the corresponding value of r_i required for calculating $A_{t\,\tau}^{(j)}$, where $j = 1, \ldots, N$.
- If

  $$A_{t\,\tau}^{(j)} < \hat{s}_t \hat{s}_\tau \hat{\rho}_{t\,\tau} + \hat{\mu}_t \hat{\mu}_\tau < A_{t\,\tau}^{(j+1)}$$

  where $j = 1, \ldots, N - 1$, then

  $$r_{t\,\tau} = 0.5(r^{(j)} + r^{(j+1)})$$

  can be taken as an approximated root of equation (3) if N is large enough.

Step 4. Estimate the coefficients of the base process.

In this step, we need to estimate $\theta_{\tau\,t}$ for $t = 1, \ldots, T-1$ and $\tau = 1, \ldots, t$. Specifically, for $t = 1$, we have $z_1 = \epsilon_1$, so no coefficient needs to be estimated.

For $t = 2$, since

$$z_2 = \theta_{11} z_1 + \epsilon_2,$$

we have

$$z_2 z_1 = \theta_{11} z_1 z_1 + \epsilon_2 z_1.$$

Therefore, by taking expectation on both sides, we get $\theta_{11} = r_{12}$.

Similarly, for $t = 3, \ldots, T$ and $\tau = t-1, \ldots, 1$, we have

$$z_t z_\tau = \theta_{1\,t-1} z_{t-1} z_\tau + \theta_{2\,t-1} z_{t-2} z_\tau + \cdots + \theta_{t-1\,t-1} z_1 z_\tau + \epsilon_t \epsilon_\tau \,.$$

Taking expectation and letting $\tau = t-1, \ldots, 1$, we get

$$\begin{aligned}
r_{t-1\,t} &= \theta_{1\,t-1} + \theta_{2\,t-1} r_{t-2\,t-1} + \cdots + \theta_{t-1\,t-1} r_{1\,t-1} \,, \\
r_{t-2\,t} &= \theta_{1\,t-1} r_{t-2\,t-1} + \theta_{2\,t-1} + \cdots + \theta_{t-1\,t-1} r_{1\,t-2} \,, \\
&\vdots \\
r_{1\,t} &= \theta_{1\,t-1} r_{1\,t-1} + \theta_{2\,t-1} r_{2\,t-1} + \cdots + \theta_{t-1\,t-1} \,.
\end{aligned}$$

In matrix form, we have

$$\begin{bmatrix} r_{t-1\,t} \\ r_{t-2\,t} \\ \vdots \\ r_{1\,t} \end{bmatrix} = \begin{bmatrix} 1 & r_{t-2\,t-1} & r_{t-3\,t-1} & \cdots & r_{1\,t-1} \\ r_{t-2\,t-1} & 1 & r_{t-3\,t-2} & \cdots & r_{1\,t-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{1\,t-1} & r_{1\,t-2} & r_{1\,t-3} & \cdots & 1 \end{bmatrix} \begin{bmatrix} \theta_{1\,t-1} \\ \theta_{2\,t-1} \\ \theta_{3\,t-1} \\ \vdots \\ \theta_{t-1\,t-1} \end{bmatrix} . \tag{4}$$

Therefore, $\theta_{\tau\,t}$ can be obtained by solving the above sequence of linear system of equations, where $t = 3, \ldots, T$ and $\tau = t-1, \ldots, 1$.

However, due to the special structure of the sequence of the linear system of equations, the values of $\theta_{\tau t}$ ($t = 1, \ldots, T-1$ and $\tau = 1, \ldots, t$) can be obtained by using the following method.

Let $\boldsymbol{\theta}_{t-1}^\top = (\theta_{1\,t-1}, \ldots, \theta_{t-1\,t-1})$, $\mathbf{r}_{t-1}^\top = (r_{t-2\,t-1}, \ldots, r_{1\,t-1})$, $R_1 = 1$ and

$$R_{t-1} = \begin{bmatrix} 1 & \mathbf{r}_{t-1}^\top \\ \mathbf{r}_{t-1} & R_{t-2} \end{bmatrix},$$

where $t \geq 3$. Then the solution to the system of equations given by (4) is given by

$$\boldsymbol{\theta}_{t-1} = R_{t-1}^{-1} \mathbf{r}_t,$$

where R_{t-1}^{-1} can be calculated recursively with details given below. Let

$$R_{t-1}^{-1} = \begin{bmatrix} a_{t-1} & \mathbf{b}_{t-1}^\top \\ \mathbf{b}_{t-1} & \mathbf{c}_{t-1} \end{bmatrix},$$

then it follows from Bernstein (2005) that

$$a_{t-1} = (1 - \mathbf{r}_{t-1}^\top R_{t-2}^{-1} \mathbf{r}_{t-1})^{-1},$$

$$\mathbf{b}_{t-1}^\top = -a_{t-1} \mathbf{r}_{t-1}^\top R_{t-2}^{-1}, \quad \mathbf{c}_{t-1} = R_{t-2}^{-1} - R_{t-2}^{-1} \mathbf{r}_{t-1} \mathbf{b}_{t-1}^\top.$$

It is easy to see that $R_1^{-1} = 1$, hence R_{t-1}^{-1} can be easily calculated for any $t \geq 3$. Therefore, $\theta_{\tau t}$ can also be obtained easily.

Step 5. Use the base process to obtain a simulated time series and then transform it to the required time series.

In this step, we simulate z_t ($t = 1, \ldots, T$) from the base process, and let $y_t = F_t^{-1}(\Phi(z_t))$ ($t = 1, \ldots, T$). Then y_t is the simulated time series with the required marginal distributions and the required autocorrelation structure.

We have implemented our method by using the Matlab program language and have used the developed software for both simulation studies and applications, the results of which are given below.

3 Simulation studies

3.1 Simulation study 1

In this simulation study we consider the following AR(3) model

$$y_t = 0.5y_{t-1} - 0.3y_{t-2} - 0.2y_{t-3} + \epsilon_t, \tag{5}$$

where $\epsilon_t \sim N(0, 0.1^2)$, which defines a stationary time series.

A time series of length 2000 was simulated from model (5). To remove the effect of initial values we only saved the last 30 values. Hence we have an "observed" time series of length $T = 30$. The above procedure was repeated 500 times. Therefore, we have $I = 500$ independent "observed" time series from model (5), each of length $T = 30$.

Once the sample autocorrelation $\hat{\rho}_{t\tau}$ and the marginal distributions of the observed time series have been obtained, the autocorrelation structure and the coefficients of the base process can then be estimated. Let $\hat{r}_{t\tau}$ and $\hat{\theta}_{ij}$ be the estimated values of $r_{t\tau}$ and θ_{ij} respectively for all possible values of t, τ, i and j. Then the estimated base process z_t is given by:

$$\begin{aligned}
&z_1 = \epsilon_1,\\
&z_2 = 0.488z_1 + \epsilon_2,\\
&z_3 = 0.79z_2 - 0.561z_1 + \epsilon_3,\\
&z_4 = 0.57z_3 - 0.35z_2 - 0.235z_1 + \epsilon_4,\\
&z_5 = 0.61z_4 - 0.373z_3 - 0.253z_2 + 0.156z_1 + \epsilon_5,\\
&\vdots\\
&z_{29} = 0.63z_{28} - 0.119z_{27} - 0.408z_{26} + \cdots + 0.088z_2 - 0.078z_1 + \epsilon_{29},\\
&z_{30} = 0.632z_{29} - 0.397z_{28} - 0.299z_{27} + \cdots + 0.067z_2 - 0.106z_1 + \epsilon_{30},
\end{aligned} \tag{6}$$

where ϵ_t ($t = 1, \ldots, 30$) are independently and normally distributed with

$$\epsilon_1 \sim N(0, 1), \qquad \epsilon_t \sim N(0, \hat{\sigma}_t^2),$$

$$\hat{\sigma}_t^2 = 1 - \hat{\theta}_{1\,t-1}\hat{r}_{t-1\,t} - \hat{\theta}_{2\,t-1}\hat{r}_{t-2\,t} - \cdots - \hat{\theta}_{t-1\,t-1}\hat{r}_{1\,t}.$$

By using the estimated base process (6) we obtained 500 simulated base time series $z_{i\,t}$ ($i = 1, 2, \ldots, 500, t = 1, 2, \ldots, 30$) and transformed them into 500 time series using $\hat{y}_{i\,t} = F_t^{-1}(\Phi(z_{i\,t}))$ ($i = 1, 2, \ldots, 500, t = 1, 2, \ldots, 30$). Note that there are no restrictions

on the number of simulated time series that can be obtained from this method. In fact we can simulate as many time series as we like.

If the simulation method performs well, then we would expect that the marginal distributions and the autocorrelation structure of the simulated time series should be similar to those of the observed time series. As the length of each time series is 30 for this simulation study, we obtained 30 marginal density functions corresponding to each time point for both observed and simulated time series. For illustration purposes, Figure 1 (a)-(d) show four plots of the estimated marginal density functions at times $t = 2, 10, 19$ and 27 respectively, where the continuous curves correspond to those obtained from the observed time series, while the dotted curves from the simulated time series. All other plots are very similar. It is clear that the simulation method reproduced the marginal distributions of the observed data with a very high quality.

Figure 1 (e) shows the plot of $\hat{\rho}_{t\,\tau} = corr(y_t, y_\tau)$ against $\bar{\rho}_{t\,\tau} = corr(\hat{y}_t, \hat{y}_\tau)$ for $t, \tau = 1, \ldots, T$, where y_t is the observed series and $\hat{y}_t$ the simulated series. It is seen that the points on the plot are very close to the straight line $y = x$, suggesting that the estimated autocorrelation structure of the simulated time series is very similar to that of the observed time series. To further quantify the difference between the two autocorrelation functions, we let

$$d = \frac{1}{T}\sqrt{\sum_{t=1}^{T}\sum_{\tau=1}^{T}\left(\hat{\rho}_{t\,\tau} - \bar{\rho}_{t\,\tau}\right)^2}\,.$$

Then d provides an average measure of the differences between two autocorrelation structures. For this simulation, we have $d = \sqrt{2.1071}/30 = 0.0484$, indicating that the difference between the two autocorrelation structures is indeed very small.

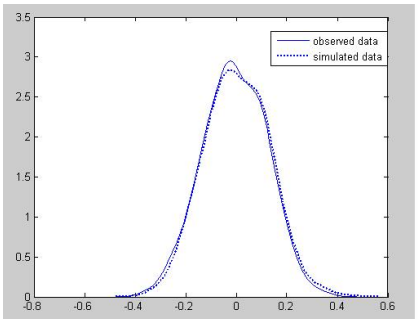
To further assess if the overall dynamics of the observed time series is preserved by the simulated realizations, we fitted an AR(3) model

$$y_t = \alpha y_{t-1} + \beta y_{t-2} + \gamma y_{t-3} + \varepsilon_t, \tag{7}$$

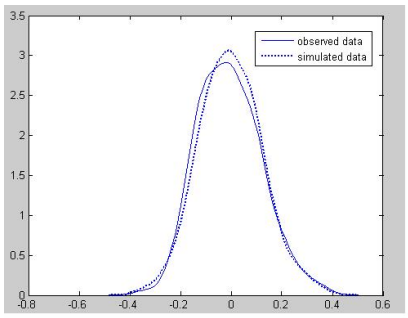
to each observed and simulated time series, where ε_t are iid $N(0, \sigma^2)$. In this simulation study, the true values of the model parameters are $\alpha = 0.5$, $\beta = -0.3$, $\gamma = -0.2$ and $\sigma = 0.1$. If the simulated time series can preserve the overall dynamics of the observed time series, then we would expect that the distribution of the estimated parameter values for the observed time series should be similar to that for the simulated time series.

By fitting an AR(3) model to each observed time series, we have estimated parameter values $\hat{\alpha}_i$, $\hat{\beta}_i$, $\hat{\gamma}_i$ and $\hat{\sigma}_i$ for $i = 1, \ldots, 500$. Figure 2 shows the corresponding probability density function plots (continuous curves) of these estimated parameter values.

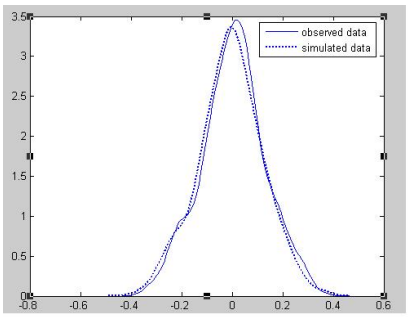
Similarly, by fitting an AR(3) to each simulated time series, we have $\tilde{\alpha}_i$, $\tilde{\beta}_i$, $\tilde{\gamma}_i$ and $\tilde{\sigma}_i$ for $i = 1, \ldots, 500$. Figure 2 also shows the corresponding histograms of the estimated parameters. Note that the vertical lines correspond to the true parameter values. It is clear that the distributions of estimated the model parameter values obtained from the observed and simulated time series respectively are very similar, suggesting that the overall dynamics of the observed time series is preserved by the simulated realizations.



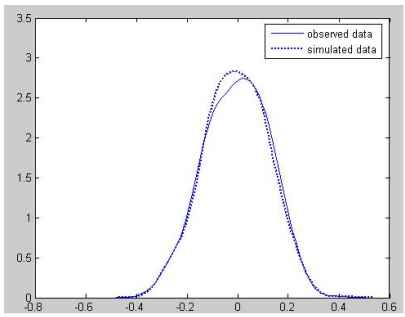
(a) t = 2



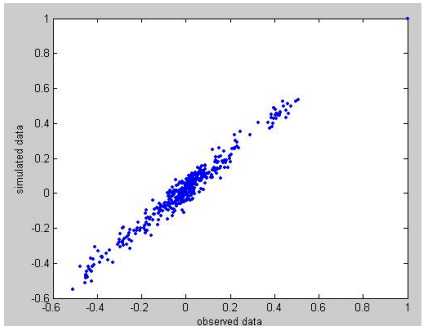
(b) t = 10



(c) t = 19



(d) t = 27



(e)

Figure 1: (a)-(d): Marginal probability density function plots of the observed (continuous curves) and the simulated (dotted curves) time series. (e) Plot of the autocorrelation coefficients of the observed time series against that of the simulated time series in Simulation study 1.

3.2 Simulation study 2

Consider a non-stationary time series defined by

$$y_t = 0.5y_{t-1} - 0.3y_{t-2} - 0.2y_{t-3} + \epsilon_t, \tag{8}$$

where $\epsilon_t \sim N(0, \sigma_t^2)$. and $\sigma_t \sim U(0.1, 0.4)$.

Similar to Simulation study 1, we generated 500 independent time series each of length 30 from model (8). By applying our method to these time series, we obtained the base process z_t given by

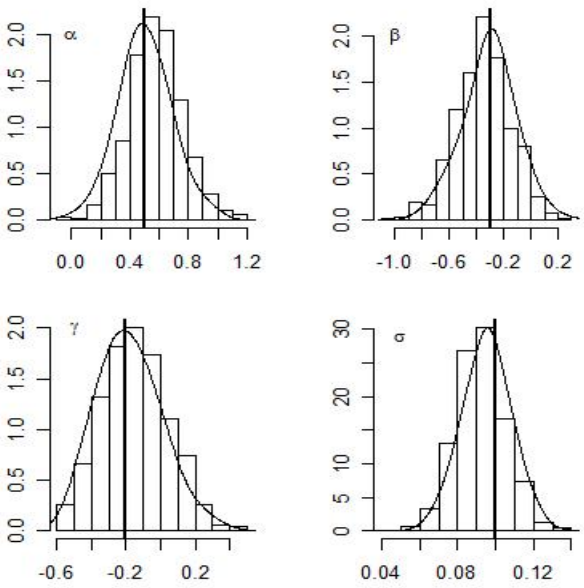
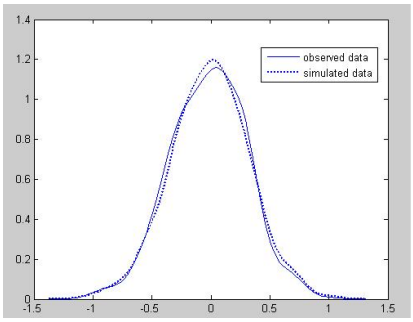


Figure 2: Distributions of the estimated model parameters in Simulation Study 1. Continuous curves: for estimated parameter values from observed series; histograms: for estimated parameter values from simulated series; vertical lines: for true parameter values.

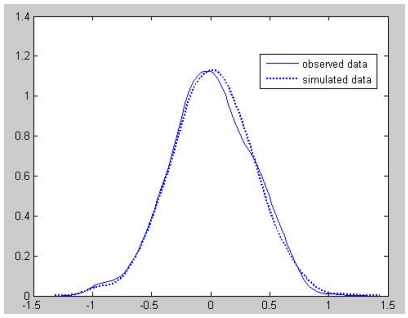
$$
\begin{aligned}
&z_1 = \epsilon_1, \\
&z_2 = 0.396z_1 + \epsilon_2, \\
&z_3 = 0.481z_2 - 0.416z_1 + \epsilon_3, \\
&z_4 = 0.448z_3 - 0.247z_2 - 0.242z_1 + \epsilon_4, \\
&z_5 = 0.504z_4 - 0.305z_3 - 0.247z_2 + 0.026z_1 + \epsilon_5, \\
&\vdots \\
&z_{29} = 0.5z_{28} - 0.326z_{27} - 0.184z_{26} + \ldots - 0.008z_3 + 0.036z_1 + \epsilon_{29}, \\
&z_{30} = 0.482z_{29} - 0.333z_{28} - 0.289z_{27} + \ldots - 0.04z_2 - 0.017z_1 + \epsilon_{30}.
\end{aligned}
$$

Hence, simulated time series can be obtained by transforming the time series generated from the base process. Again 500 simulated time series each of length 30 were obtained. Figure 3 (a)-(d) show the marginal density function plots at times $t = 2$, 9, 18 and 27 for illustration purposes, where the continuous curves correspond to those obtained from the observed time series, while the dotted curves obtained from the simulated time series. Figure 3 (e) shows the plot of $\hat{\rho}_{t\tau}$ against $\bar{\rho}_{t\tau}$ for this simulation study, indicating good performance of the developed method for this simulation study. Furthermore, the difference between the two autocorrelation functions is $d = \sqrt{1.3934}/30 = 0.0393$, which is also very small.

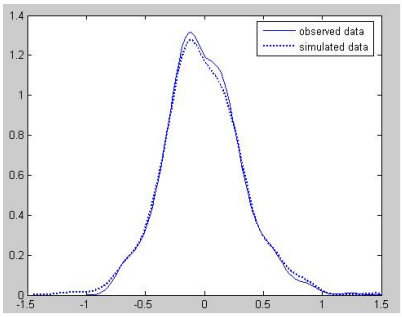
Different from model (5), the variance of ϵ_t in model (8) is not a constant. However, to assess if the overall dynamics of the observed time series is preserved by the simulated realizations, we may still fit model (7) to the observed and simulated time series in this



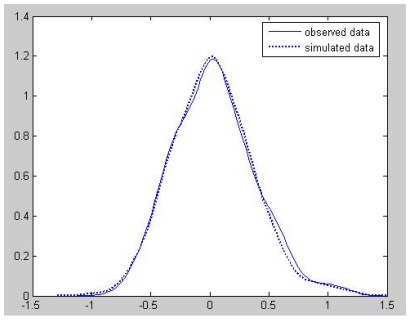
(a) t = 2



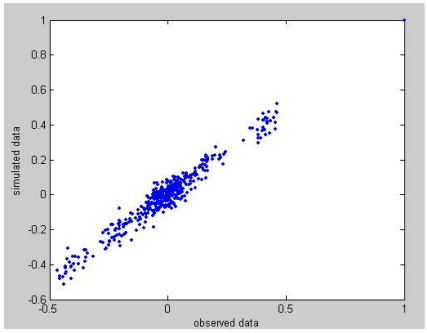
(b) t = 9



(c) t = 18



(d) t = 27



(e)

Figure 3: (a)-(d): Marginal probability density function plots of the observed (continuous curves) and the simulated (dotted curves) time series. (e) Plot of the autocorrelation coefficients of the observed time series against that of the simulated time series in Simulation study 2.

---

simulation study. This is because model (7) should perform similar for both observed and simulated time series if our simulation method works well. Figure 4 confirms that indeed the overall dynamics of both observed and simulated time series is also very similar in this simulation study.

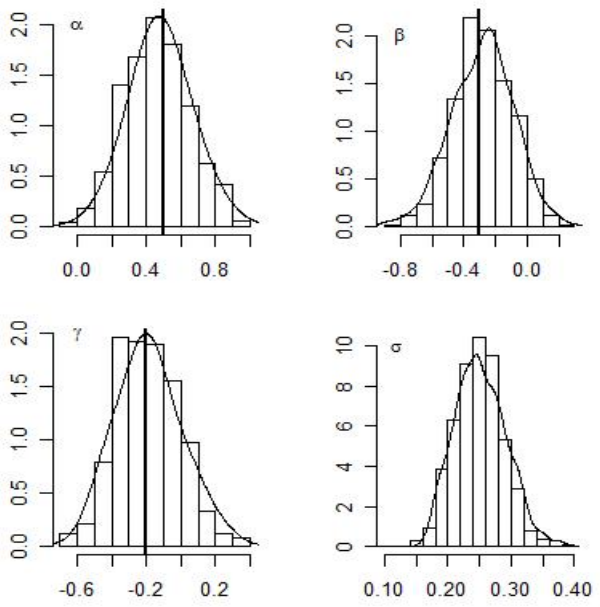


Figure 4: Distributions of the estimated model parameters in Simulation Study 2. Continuous curves: for estimated parameter values from observed series; histograms: for estimated parameter values from simulated series; vertical lines: for true parameter values.

3.3 Simulation study 3

Now consider another non-stationary process defined by

$$y_t = \sin(y_{t-1}) + \epsilon_t \tag{9}$$

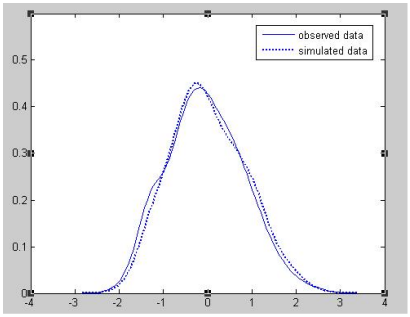
where $\epsilon_t \sim N(0, 0.6^2)$.

In this simulation study, 300 independent time series each of size 20 were generated from model (9). By applying our simulation method, we obtained the estimated base process z_t given by

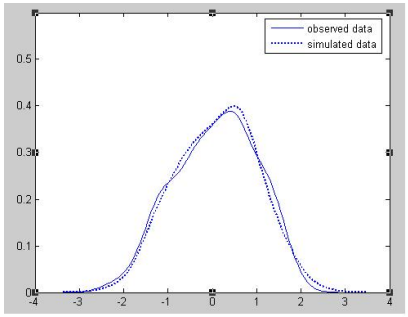
$$\begin{aligned}
&z_1 = \epsilon_1,\\
&z_2 = 0.704z_1 + \epsilon_2,\\
&z_3 = 0.736z_2 + 0.011z_1 + \epsilon_3,\\
&z_4 = 0.746z_3 + 0.004z_2 - 0.012z_1 + \epsilon_4,\\
&z_5 = 0.737z_4 + 0.071z_3 - 0.098z_2 + 0.08z_1 + \epsilon_5,\\
&\vdots\\
&z_{19} = 0.572z_{18} + 0.223z_{17} - 0.151z_{16} + \ldots - 0.046z_2 - 0.05z_1 + \epsilon_{19},\\
&z_{20} = 0.777z_{19} - 0.021z_{18} - 0.111z_{17} + \ldots + 0.069z_2 - 0.068z_1 + \epsilon_{20}.
\end{aligned}$$

Figure 5 (a)-(d) show the marginal density function plots corresponding to times $t =$ 1, 5, 9 and 17 respectively, where the continuous curves correspond to those obtained

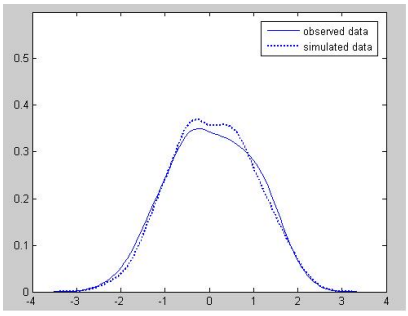
from the observed time series, while the dotted curves from the simulated time series. Figure 5 (e) compares the difference between autocorrelation structures of the simulated and observed time series, which also suggests that a very good agreement between them has been achieved. Indeed, in this case we have $d = \sqrt{0.896}/20 = 0.0473$, which further confirms that the method also works well in this simulation study.



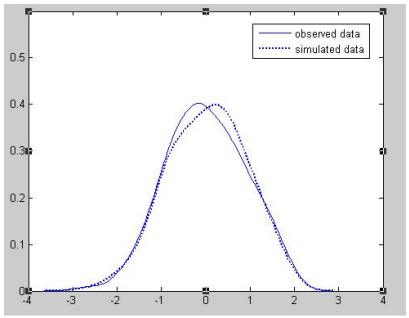
(a) t = 1



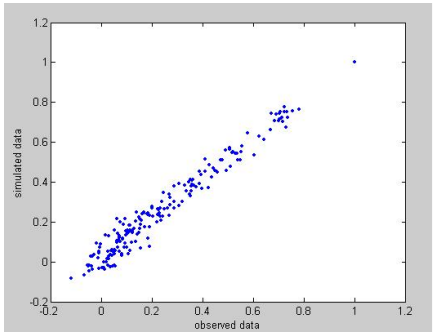
(b) t = 5



(c) t = 9



(d) t = 17



(e)

Figure 5: (a)-(d): Marginal probability density function plots of the observed (continuous curves) and the simulated (dotted curves) time series. (e) Plot of the autocorrelation coefficients of the observed time series against that of the simulated time series in Simulation study 3.

To assess if the overall dynamics of the observed time series is preserved by the simulated realizations, we may fit the following model to both observed and simulated time series from this simulation study:

$$y_t = \alpha \sin(y_{t-1}) + \varepsilon_t,$$

where ε_t are iid with zero mean and constant variance σ^2. The true α value is 1 and the true σ value is 0.6. The results are shown in Figure 6, which again suggests that the simulated time series can preserve the overall dynamics of the observed time series.

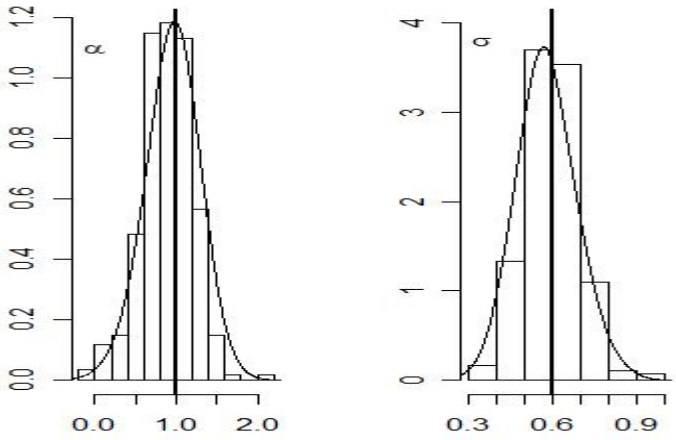


Figure 6: Distributions of the estimated model parameters in Simulation Study 3. Continuous curves: for estimated parameter values from observed series; histograms: for estimated parameter values from simulated series; vertical lines: for true parameter values.

4 Applications

We now apply our method to a real data set, which consists of $I = 258$ size series of cocoon filament, each of length $T = 12$. The first two rows of Figure 7 show randomly selected four observed series (others are very similar). It is seen that the time series plots do suggest that they are non-stationary.

By applying our method to the data we obtained the base process given below:

$$
\begin{aligned}
z_1 &= \epsilon_1, \\
z_2 &= 0.81z_1 + \epsilon_2, \\
z_3 &= 1.101z_2 - 0.34z_1 + \epsilon_3, \\
z_4 &= 1.128z_3 - 0.338z_2 + 0.009z_1 + \epsilon_4, \\
z_5 &= 0.937z_4 - 0.004z_3 - 0.076z_2 - 0.084z_1 + \epsilon_5, \\
z_6 &= 0.9z_5 + 0.11z_4 - 0.133z_3 - 0.16z_2 + 0.085z_1 + \epsilon_6, \\
z_7 &= 0.809z_6 + 0.161z_5 - 0.073z_4 - 0.033z_3 - 0.142z_2 + 0.096z_1 + \epsilon_7, \\
z_8 &= 0.883z_7 + 0.174z_6 - 0.125z_5 - 0.195z_4 - 0.003z_3 + 0.092z_2 - 0.065z_1 + \epsilon_8,
\end{aligned}
$$

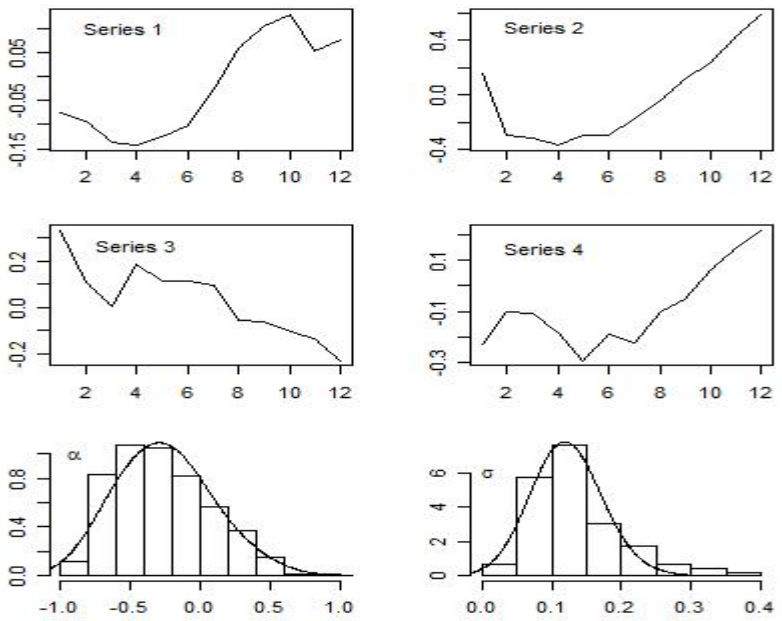


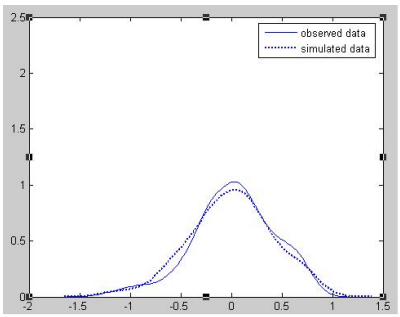
Figure 7: First two rows: Time series plots of four randomly selected size series of cocoon filament. Last row: Distributions of the estimated parameters of model (10) for size series of cocoon filament.

$z_9 = 0.899z_8 + 0.203z_7 - 0.291z_6 + 0.056z_5 - 0.17z_4 - 0.115z_3 + 0.122z_2$
$-0.129z_1 + \epsilon_9,$
$z_{10} = 0.868z_9 - 0.018z_8 - 0.028z_7 - 0.016z_6 - 0.101z_5 - 0.002z_4 - 0.082z_3$
$+0.04z_2 - 0.091z_1 + \epsilon_{10},$
$z_{11} = 0.82z_{10} + 0.052z_9 + 0.095z_8 - 0.263z_7 - 0.14z_6 + 0.065z_5 + 0.086z_4$
$-0.067z_3 - 0.109z_2 - 0.014z_1 + \epsilon_{11},$
$z_{12} = 0.956z_{11} - 0.2z_{10} + 0.067z_9 - 0.274z_8 + 0.119z_7 - 0.026z_6 - 0.016z_5$
$-0.091z_4 + 0.014z_3 - 0.023z_2 - 0.068z_1 + \epsilon_{12}.$

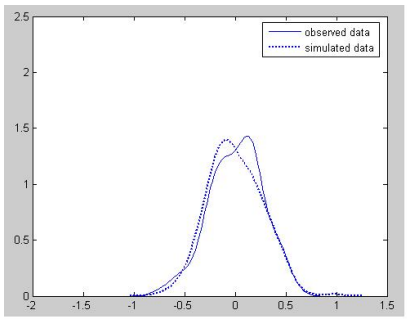
Hence, we generated 258 simulated size series, based on which, marginal probability density functions at different time points can be obtained. For illustration purposes, Figure 8 (a)-(d) show the marginal probability density function plots at times $t = 1$, 4, 7 and 11 respectively, where the continuous curves correspond to those obtained from the observed time series, while the dotted curves from the simulated time series. Figure 8 (a)-(d) also clearly show the non-stationarity of the size series. Figure 8 (e) further compared the autocorrelation structures of the simulated and the observed time series. The differences between them is measured by $d = \sqrt{0.349}/12 = 0.0492$. All the results show that the simulated time series are in a very good agreement with the observed size series of cocoon filament.

It is worth mentioning that the value of N controls the accuracy of the estimated value of $r_{t\tau}$. For example, in this application, we simply let $N = I = 258$, leading to $r^{(j+1)} - r^{(j)} = 2/(258 + 1) = 0.00772$. Hence the difference between $r_{t\tau}$ and its estimated value

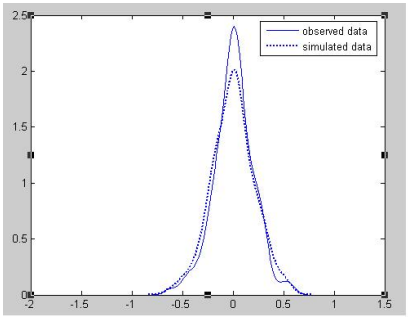
$0.5 * (r^{(j+1)} + r^{(j)})$ is less than $0.5 * 0.00772 \approx 0.00386$. We have found that an accuracy at this level leads to good simulated time series. However, the optimal choice of N requires further investigation in the future.



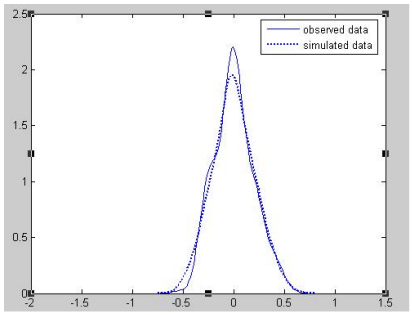
(a) t = 1



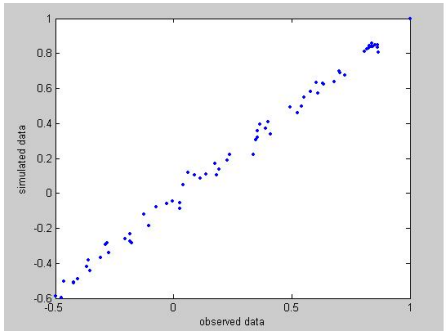
(b) t = 4



(c) t = 7



(d) t = 11



(e)

Figure 8: (a)-(d): Marginal probability density function plots of the observed (continuous curves) and the simulated (dotted curves) time series. (e) Plot of the autocorrelation coefficients of the observed time series against that of the simulated time series for the real data set.

Although we are unable to carry out further analysis that could be done by the researchers in the raw reeling silk industry, we could assess if the overall dynamics of the observed size series of cocoon filament is preserved by the simulated realizations as we did in the simulation studies. For this application, we used the ARIMA(1, 2, 0) model

$$u_t = \alpha u_{t-1} + \varepsilon_t, \quad u_t = (1 - B)^2 y_t, \tag{10}$$

where y_t is the value of a size series at time t, B is the backward shift operator such that $By_t = y_{t-1}$, and ε_t are iid with zero mean and constant variance σ^2. We have found that a

second order differencing of the size series is necessary in order to fit an ARIMA(1, 2, 0) model to each of the observed and simulated size series. The last row of Figure 7 shows that the distribution of the estimated α values from both observed and simulated size series are very similar, which is also true for σ. These results suggest that the simulated size series have the same autocorrelation structure and the same marginal distribution as those of the observed size series. Furthermore, the simulated size series can also preserve the overall dynamics of the observed size series from a statistical modelling point of view.

5 Further comments and conclusions

We have developed a simulation method for non-stationary time series of finite length. The developed simulation method guarantees that the simulated time series are in a very good agreement with the observed time series with respect to the marginal distributions and the autocorrelation structure. We have demonstrated that our method can provide high quality simulated time series for researchers in raw reeling silk industry. We also expect that the developed method is of great importance in many other areas. For example, in medical research it can be very expensive to follow a large number of patients for a period of time, and in coastal engineering it can also be very expensive to collect sea condition data at many locations. Our method enables the researchers to carry out further analysis based both on the simulated and on the observed time series, leading to significant financial savings.

We assessed if the overall dynamics of the observed time series is preserved by the simulated time series by fitting a statistical model to both observed and simulated time series. For the cases considered in this paper good results have been obtained. However, it is worth mentioning that the developed methodology does not guarantee the joint marginal distributions. Therefore, some dynamic features of an observed non-stationary time series may not be preserved.

For example, let us consider a self-exciting threshold autoregressive time series defined by

$$y_t = \sum_{k=1}^{K} \left(a_{k0} + \sum_{j=1}^{p} a_{kj} y_{t-j} + \epsilon_{kt} \right) I_{[y_{t-d} \in \Omega_k]},$$

where $\Omega_k = [r_{k-1}, r_k)$, and $-\infty < r_0 < r_1 < \cdots < r_K < \infty$ are threshold values. As the joint marginal distributions of the process can be very complicated due to the unknown threshold values involved, we would expect that some important dynamic features in high dimensions may not be preserved by the developed method.

We feel that it can be very difficult to cover all dynamic features of an observed non-stationary time series within the framework developed in this paper. Further research is certainly required in the future.

It is of great interest to investigate the effects of the simulated time series on the further analysis required by the researchers in the raw reeling silk industry or other areas, but this is beyond the scope of this paper.

Acknowledgement

This research was supported by NNSF of China (Grant No. 10801104) and NSF of Jiangsu Province (Grant No. BK2012612). We would also like to express our sincere thanks to Professor Wanchun Fei, College of Textile and Clothing Engineering, Soochow University, China, for providing us with the real data set.

We thank the referees for their very constructive comments and suggestions, which have greatly enhanced the quality and the presentation of the paper.

Appendix

Proof of Theorem 1: We use the induction rule to prove the theorem. It is true for $t = 1$. Suppose it is also true up to time $t-1$, that is, $z_\tau \sim N(0,1)$ $(\tau = 1, \ldots, t-1)$, we need to show that $z_t \sim N(0,1)$.

It follows from (1) that z_t is a linear combination of normal random variables, hence z_t is normally distributed. Furthermore,

$$E(z_t) = \theta_{1\,t-1}E(z_{t-1}) + \theta_{2\,t-1}E(z_{t-2}) + \cdots + \theta_{t-1\,t-1}E(z_1) + E(\epsilon_t) = 0,$$

and

$$\begin{aligned}
&var(z_t) = E(z_t z_t) \\
&= \theta_{1\,t-1}E(z_{t-1}z_t) + \theta_{2\,t-1}E(z_{t-2}z_t) + \cdots + \theta_{t-1\,t-1}E(z_1 z_t) + E(\epsilon_t z_t) \\
&= \theta_{1\,t-1}r_{t-1\,t} + \theta_{2\,t-1}r_{t-2\,t} + \cdots + \theta_{t-1\,t-1}r_{1\,t} + var(\epsilon_t) \\
&= \theta_{1\,t-1}r_{t-1\,t} + \theta_{2\,t-1}r_{t-2\,t} + \cdots + \theta_{t-1\,t-1}r_{1\,t} \\
&\quad +(1 - \theta_{1\,t-1}r_{t-1\,t} - \theta_{2\,t-1}r_{t-2\,t} - \cdots - \theta_{t-1\,t-1}r_{1\,t}) \\
&= 1.
\end{aligned}$$

It follows from the induction rule that $z_t \sim N(0,1)$ for any t.

□

Proof of Theorem 2: First note that by following the lines of their work, the Propositions 1 and 2 of Cario and Nelson (1996) also hold in our case.

Then note that the Lemma A.1. of Cario and Nelson (1996) holds when replacing one nondecreasing function g by two non-decreasing function g_1 and g_2. This is because the Proposition 1 of Rubinstein et al. (1985) holds for any two non-decreasing functions. So by taking $g_1 = F_t^{-1}(\Phi(\cdot))$ and $g_2 = F_\tau^{-1}(\Phi(\cdot))$, we have that $E(y_t y_\tau)$ is nondecreasing.

Finally, replacing $Y_1 = F_Y^{-1}(\Phi(Z_1))$ by $y_t = F_t^{-1}(\Phi(z_t))$ and $Y_2 = F_Y^{-1}(\Phi(Z_2))$ by $y_\tau = F_\tau^{-1}(\Phi(z_\tau))$ respectively in the proof of Lemma A.2. of Cario and Nelson (1996), we have that $E(y_t y_\tau)$ is also continuous on $-1 \leq r_{t\tau} \leq 1$. □

References

[1] Barone, Piero (1987). A method for generating independent realizations of a multivariate normal stationary and invertible ARMA(p, q) process. *Journal of Time Series Analysis*, 8, 125-130.

[2] Bernstein, Dennis (2005). Matrix Mathematics. Princeton University Press.

[3] Biller, Bahar and Nelson, Barry L. (2003). Modeling and generating multivariate time-series input processes using a vector autoregressive technique. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 13, 211-237.

[4] Biller, Bahar and Nelson, Barry L. (2008). Evaluation of the ARTAFIT method for fitting time-series input processes for simulation. *INFORMS Journal on Computing*, 20, 485-498.

[5] Cai, Yuzhi (2011). Multivariate time series simulation. *Journal of Time Series Analysis*, 32, 566-579.

[6] Cai, Yuzhi, Gouldby, Ben, Hawkes, Peter and Dunning, Paul (2008). Statistical Simulation of Flood Variables: Incorporating Short-Term Sequencing. Journal of Flood Risk Management. Vol.1, 1-10.

[7] Cario, Marne C. and Nelson, Barry L. (1996), Autoregressive to anything: Time series input processes for simulation. *Operations Research Letters*, 19: 51-58.

[8] Cario, Marne C. and Nelson, Barry L. (1998), Numerical methods for fitting and simulating autoregressive-to-anything processes. *Journal on Computing*, 10, 72-81.

[9] Deler, Bahar and Nelson, Barry L. (2001). Modeling and generating multivariate time series with arbitrary marginals and autocorrelation structures. *Proceedings of the 2001 Winter Simulation Conference. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers*, 275-282.

[10] Fei, Wanchun and Bay, Lun (2005). Auto-regressive models of non-stationary time series with finite length. *Tsinghua Science and Technology*, 10, 162-168.

[11] Fei, Wanchun and Bay, Lun (2009). Time-varying parameter auto-regressive models for autocovariance nonstationary time series. *Science in China Series A: Mathematics*, 52, 577-584.

[12] Li, Jenny and Winker, Peter (2003). Time Series Simulation with Quasi Monte Carlo Methods. *Computational Economics*, 21 , 23-43.

[13] Shea, Brain L. (1988). A Note on the generation of independent realizations of a vector autoregressive moving-average process. *Journal of Time Series Analysis*, 9, 403-410.

[14] Rubinstein, R.Y., Samorodnitsky, G. and Shaked, M. (1985). Antithetic variates, multivariate dependence and simulation of stochastic systems. *Management Science*, 31: 66-77.