COMMUNICATIONS IN STATISTICS - SIMULATION AND COMPUTATION®
https://doi.org/10.1080/03610918.2019.1602647

Model diagnostic procedures for copula-based Markov chain models for statistical process control

Xin-Wei Huang and Takeshi Emura

Graduate Institute of Statistics, National Central University, Taoyuan, Taiwan

ABSTRACT

Investigating serial dependence is an important step in statistical process control (SPC). One recent approach is to fit a copula-based Markov chain model to perform SPC, which provides an attractive alternative to the traditional AR1 model. However, methodologies for model diagnostic have not been considered. In this paper, we develop two different approaches for model diagnostic procedures for copula-based Markov chain models. The first approach employs a formal test based on the Kolmogorov-Smirnov or the Cramér-von Mises statistics with aid of a parametric bootstrap. The second approach employs the second-order Markov chain model to examine the Markov property in the model. This second approach itself is a new SPC method. We made all the computing methodologies available in the R *Copula.Markov* package, and check their performance by simulations. We analyze three datasets for illustration.

ARTICLE HISTORY
Received 4 February 2019
Accepted 28 March 2019

KEYWORDS
Control chart; Copulas; goodness-of-fit tests; Markov chain; serial dependence; Statistical process control; Time series

MATHEMATICS SUBJECT CLASSIFICATION
62M10; 62H12; 62H20; 60J20

1. Introduction

Observed data collected in daily manufacturing process are often dependent in the sense that the present sampling condition depends on the past ones. Thus, modeling dependence in the data plays a crucial role in statistical process control (SPC) (Montgomery 2009).

Let $\{Y_t : t = 1, ..., n\}$ be data collected on n different time points. In some cases, an unusually high (or low) value of Y_{t-1} may influence the next value of Y_t (Bisgaard and Kulahci 2007). While the major goal of SPC is to monitor the marginal process parameters (mean and SD of Y_t), the dependence parameter (e.g. $\text{cor}(Y_{t-1}, Y_t)$) largely influences long-term performance of SPC, such as the average run length (ARL).

A solid overview of serial dependence models in SPC is found in Wieringa (1999) and Knoth and Schmid (2004), while a concise review is seen in Box and Narasimhan (2010). The literature focuses on the first order (Markov) models, including a first order autoregressive AR(1), a first order moving average MA(1), and a first order integrated moving average IMA(1) or IMA(1,1). These traditional models can only deal with linear dependence between two observations. For instance, the AR(1) model applies a linear structure

CONTACT Takeshi Emura takeshiemura@gmail.com Graduate Institute of Statistics, National Central University, Taoyuan, Taiwan.

$$Y_t = \xi + \rho Y_{t-1} + \varepsilon_t, \quad t = 1, ..., n,$$

where $\varepsilon_t \sim^{iid} N(0, \tau^2)$, $-1<\rho<1$, and $\tau^2>0$.

Long and Emura (2014) considered a copula-based Markov chain model to perform SPC for serially correlated data. In their model, serial dependence between a pair of consecutive observations (Y_{t-1}, Y_t) is modeled as

$$\Pr(Y_t \le y_t, Y_{t-1} \le y_{t-1}) = C\big(G(y_t), G(y_{t-1})\big), \tag{1}$$

where $C : [0,1]^2 \to [0,1]$ is a copula (Nelsen 2006) and $G(y) = \Pr(Y_t \le y)$ is the marginal (stationary) distribution function. Note that the model (1) itself was originally proposed by Darsow, Nguyen, and Olsen (1992), and subsequently applied to different statistical problems by Joe (1997), Chen and Fan (2006); Abegaz and Naik-Nimbalkar (2008); Domma, Giordano, and Perri (2009); Huang, Chen, and Emura (2019); and Erkal Sonmez and Baray (2019). The computer tools were only recently available through the R package *Copula.Markov* (Emura, Long, and Sun 2017). The package has been applied to a number of recent studies (Kim and Baik 2018; Kim, Baik, and Reller 2018; Kim, Baik, and Reller 2019; Sun, Lee, and Emura 2018).

Applying the package to the data $\{Y_t : t = 1, ..., n\}$, one can estimate process parameter (μ, σ), where $\mu = E(Y_t)$ and $\sigma = \sqrt{\mathrm{Var}(Y_t)}$, as well as a copula parameter. Since the models of Long and Emura (2014) and Emura, Long, and Sun (2017) are fully parametric, the package requires strong distributional assumptions of the normal marginal model $G(y) = \Phi\{(y-\mu)/\sigma\}$, where Φ is the distribution function of $N(0,1)$, and a copula function (Clayton copula or Joe copula). The effect of violating the normality assumption for independent data has been noticed by Albers and Kallenberg (2007). Under the copula-based model for correlated data, the model assumptions are even more stringent. We have three model assumptions to be addressed:

i. *Markov property*: $\Pr(Y_t \le y_t | Y_{t-1} = y_{t-1}, Y_{t-1} = y_{t-2}, ...) = \Pr(Y_t \le y_t | Y_{t-1} = y_{t-1})\ \forall t$
ii. *Marginal distribution*: $G(y) = \Phi\{(y-\mu)/\sigma\}\ \exists(\mu, \sigma)$
iii. *Copula form*: $\Pr(Y_t \le y_t, Y_{t-1} \le y_{t-1}) = C_\alpha(G(y_t), G(y_{t-1}))\ \exists\alpha$

None of these assumptions were examined when fitting the data to the copula-based Markov chain model since the model diagnostic methods are unavailable in software packages.

Therefore, the main goal of this paper is to present model diagnostic procedures to examine (i)–(iii). To examine (ii), we propose significance tests based on the Kolmogorov-Smirnov and the Cramér-von Mises statistics with aid of a parametric bootstrap. This examine (i), we propose a model comparison approach with the 2^{nd}-order Markov chain model. As a byproduct, we also propose a new SPC method under the 2^{nd}-order Markov chain model. We propose to check (iii) by comparing the goodness-of-fit between the Clayton and Joe copulas (Joe 1993), and chose the better one to perform SPC. We made all the computing codes available a user-friendly manner in the R *Copula.Markov* package. We illustrate the proposed model diagnostic methods through three datasets needing SPC methods with serial dependence.

The paper is organized as follows. Section 2 reviews existing methods. Section 3 proposes goodness-of-fit tests. Section 4 proposes SPC under the 2^{nd}-order model, which can be used for a model comparison. Section 5 analyses three datasets. Section 6 concludes. Appendices A and B provide the definitions of the R functions in the R package. Appendix C provides the derivations of the likelihood function. Appendix D provides the R codes for the data analyses.

2. Parameter estimation and control chart

In this section, we review parameter estimation and SPC methods under a copula-based Markov chain model as previously proposed by Long and Emura (2014) and Emura, Long, and Sun (2017).

For observations $\{Y_t : t = 1, ..., n\}$, we assume the Markov property

$$\Pr\left(Y_t \le y_t | Y_{t-1} = y_{t-1}, Y_{t-1} = y_{t-2}, ...\right) = \Pr\left(Y_t \le y_t | Y_{t-1} = y_{t-1}\right) \ \forall t,$$

and a bivariate copula Markov model

$$\Pr(Y_t \le y_t, Y_{t-1} \le y_{t-1}) = C_\alpha\left(G(y_t), G(y_{t-1})\right), \tag{2}$$

where C_α is a copula (Nelsen 2006), $\alpha \in \mathbb{R}$ is a dependence parameter, and

$$G(y) = \Phi\{(y-\mu)/\sigma\}. \tag{3}$$

We have marginal mean $\mu = E(Y_t)$ and SD $\sigma = \sqrt{\mathrm{Var}(Y_t)}$. The reason of choosing the normal margins is due to its remarkable popularity of the three-sigma rule of $\mu \pm 3\sigma$ in SPC.

We mainly focus on the one-parameter Clayton copula defined as:

$$C_\alpha(u_1, u_2) = \left({u_1}^{-\alpha} + {u_2}^{-\alpha} - 1\right)^{-1/\alpha},$$

where $\alpha > 0$ is related to Kendall's tau between Y_{t-1} and Y_t, through $\tau = \alpha/(\alpha + 2)$. Our choice of the Clayton copula is due to its popularity. Many statistical models and software packages in biostatistics focus on the Clayton copula due to its ease of conducting simulation (e.g. Rotolo, Legrand, and Van Keilegom 2013), estimation (e.g. Emura, Long, and Sun 2017; Emura et al. 2017; Rotolo, Paoletti, and Michiels 2018; Emura, Matsui, and Rondeau 2019), feature selection (Emura and Chen 2016; Emura, Matsui, and Chen 2019) and prediction (e.g. Emura et al. 2018). The simplicity and usefulness of the Clayton copula are also true in SPC, but its goodness-of-fit to real data is not always true. The issue of the goodness-of-fit shall be examined in details.

Under the models (2) and (3), the log-likelihood function given $\{Y_t : t = 1, ..., n\}$ is

$$\ell(\mu, \sigma, \alpha) = \frac{1}{n}\sum_{t=1}^{n} \log\left\{\frac{1}{\sigma}\varphi\left(\frac{Y_t - \mu}{\sigma}\right)\right\} + \frac{1}{n}\sum_{t=2}^{n} \log c_\alpha\left\{\Phi\left(\frac{Y_{t-1} - \mu}{\sigma}\right), \Phi\left(\frac{Y_t - \mu}{\sigma}\right)\right\},$$

where $c_\alpha(u_1, u_2) = \partial^2 C_\alpha(u_1, u_2)/\partial u_1 \partial u_2$. For any chosen copula, the MLE that maximizes the preceding formula is denoted by $(\hat{\mu}, \hat{\sigma}, \hat{\alpha})$.

The *Copula.Markov* package provides two options for copulas, the Clayton copula via `Clayton.Markov.MLE()` and Joe copula via `Joe.Markov.MLE()`. Note that the Clayton copula has the lower tail dependence while the Joe copula has the upper tail

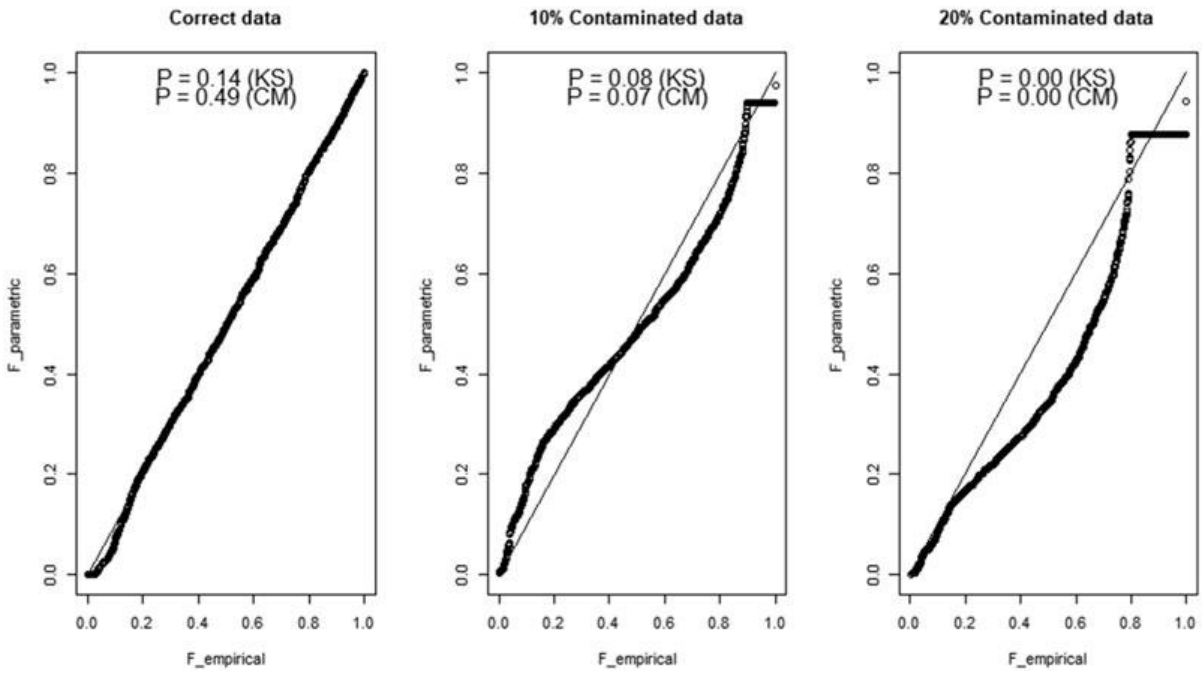

Figure 1. Diagnostic plots made by the correct data (left hand) and 10% contaminated data (center), and 20% contaminated data (right hand). We generated the correct data from the Clayton model with the normal margins under $\mu = 1$, $\sigma = 1$, and $\alpha = 2$ ($\tau = 0.5$). For the contaminated data, we randomly replace 10% (or 20%) of the data by the outlier $\mu + 3\sigma = 4$.

dependence. Hence these two copulas capture quite different dependence structures and supplement each other in modeling serial dependence. Comparison of the log-likelihood values between the Clayton and Joe copulas leads to a very simple but effective strategy for model selection.

Control charts provide a tool to detect out-of-control signals in observations $\{Y_t : t = 1, ..., n\}$. The three-sigma control chart consists of the center $\mu = E(Y_t)$ and the control limits $\mu \pm 3\sigma$, where $\sigma = \sqrt{\mathrm{Var}(Y_t)}$. If the parameters (μ, σ) are unknown, one can use the MLE to obtain the estimators of LCL and UCL, as $\hat{\mu} - 3\hat{\sigma}$ and $\hat{\mu} + 3\hat{\sigma}$, respectively. Out-of-control points are detected by Y_t>UCL, or Y_t<LCL. Control charts usually display the plot of $\{Y_t : t = 1, ..., n\}$ together with the control limits. The *Copula.Markov* package provides functions `Clayton.Markov.MLE()` and Joe copulas via `Joe.Markov.MLE()` to draw control charts.

3. Goodness-of-fit test

Since the validity of the MLE and control limits relies heavily on the model assumptions, we propose a goodness-of-fit test procedure. We are interested in testing a hypothesis

$$H_0 : \Pr(Y_t \le y) = \Phi\left(\frac{y-\mu}{\sigma}\right) \quad \text{for} \quad \exists(\mu, \sigma),$$

against an alternative hypothesis

$$H_1 : \Pr(Y_t \le y) \ne \Phi\left(\frac{y-\mu}{\sigma}\right) \quad \text{for} \quad \forall(\mu, \sigma).$$

Let $G_n(y) = \sum_{t=1}^{n} \mathbf{I}\{Y_t \le y\}/n$ be the empirical distribution function. If the model is correct, the parametric estimator $\Phi\{(y-\hat{\mu})/\hat{\sigma}\}$ and the nonparametric estimator $G_n(y)$ converges to the true value (Chen and Fan 2006; Long and Emura 2014). If the model is wrong, the two estimators converge to different values. Thus, we propose a Kolmogorov-Smirnov statistic

$$K = \sup\left|G_n(y) - \Phi\left(\frac{y-\hat{\mu}}{\hat{\sigma}}\right)\right| ,$$

and a Cramér-von Mises type statistic

$$C = \int n\left\{G_n(y) - \Phi\left(\frac{y-\hat{\mu}}{\hat{\sigma}}\right)\right\}^2 dG_n(y) = \sum_j \left\{G_n(y_j) - \Phi\left(\frac{y_j-\hat{\mu}}{\hat{\sigma}}\right)\right\}^2,$$

to detect the departure of the model from the underlying model.

We suggest a parametric bootstrap method to obtain the P-value of the tests:

The goodness-of-fit test with parametric bootstrap

Step 1: Generate Markov time series $\{Y_t^{(b)} : t = 1, ..., n\}$ under H_0 with estimated parameters $(\hat{\mu}, \hat{\sigma}, \hat{\alpha})$ for each $b = 1, 2, ..., B$.

Step 2: Compute the MLE $(\hat{\mu}^{(b)}, \hat{\sigma}^{(b)}, \hat{\alpha}^{(b)})$, the parametric estimator $\Phi\left(\frac{y-\hat{\mu}^{(b)}}{\hat{\sigma}^{(b)}}\right)$, and the nonparametric estimator $G_n^{(b)}(y)$ from the data $\{Y_t^{(b)} : t = 1, ..., n\}$. Then, compute the corresponding statistic $K^{(b)}$ or $C^{(b)}$ for each $b = 1, 2, ..., B$.

Step 3: The P-value of the test is calculated as $\sum_{b=1}^{B} \mathbf{I}(K^{*(b)} \ge K)/B$ or $\sum_{b=1}^{B} \mathbf{I}(C^{*(b)} \ge C)/B$.

Reject H_0 if the P-value is less than a specified significance level P, Otherwise, accept H_0.

We suggest $B = 500$ as the number of the bootstrap replications.

In conjunction with the test results, a graphical diagnostic procedure is useful by plotting $\Phi(\frac{Y_t-\hat{\mu}}{\hat{\sigma}})$ against $G_n(Y_t)$. If the plot bends away from the diagonal line, this indicates evidence that the fitted model is not a good choice. Figure 1 shows three plots, one for correct data, and other two for contaminated data. We see that the plot almost perfectly lines on the diagonal for the correct data while the plots bend away from the diagonal line for the contaminated data.

We implemented computation of the goodness-of-fit tests as well as the diagnostic plot under the Clayton copula (by `Clayton.Markov.GOF`) and Joe copula (by `Joe.Markov.GOF`).

We conducted simulations to examine the type I error rates and power for the proposed goodness-of-fit test. First, we generated data from the Clayton model with the normal margins under $\mu = 1$, $\sigma = 1$, and $\alpha = 2$ ($\tau = 0.5$). For each data generated, we

Table 1. The rejection rates of the goodness-of-fit tests based on 200 repetitions under $\mu = 1$, $\sigma = 1$, and $\alpha = 2$; K=Kolmogorov-Smirnov statistic; C=Cramér-von Mises statistic.

		Without outliers		Outliners with rate = 10% size=$\mu + 3\sigma$		Outliners with rate = 10% size=$\mu + 6\sigma$		Outliners with rate = 20% size=$\mu + 3\sigma$	
Sample size	Nominal Significance level P	K	C	K	C	K	C	K	C
$n = 300$	0.01	0.01	0.00	0.00	0.00	0.56	0.53	0.27	0.27
	0.05	0.03	0.04	0.00	0.01	0.93	0.89	0.86	0.88
	0.10	0.10	0.10	0.17	0.15	0.98	0.97	0.92	0.92
$n = 600$	0.01	0.00	0.00	0.02	0.01	0.87	0.88	0.74	0.81
	0.05	0.03	0.01	0.21	0.19	0.99	1.00	0.97	0.93
	0.10	0.12	0.14	0.41	0.53	1.00	1.00	0.99	0.94
$n = 1000$	0.01	0.01	0.00	0.08	0.10	0.98	0.98	0.92	0.92
	0.05	0.04	0.04	0.54	0.66	1.00	1.00	0.99	0.95
	0.10	0.10	0.10	0.84	0.91	1.00	1.00	1.00	0.97

performed the parametric bootstrap tests, and then examined the number of rejections among 200 repetitions (empirical rejection rates) under three nominal significance levels, $P = 0.01$, 0.05, and 0.10.

Table 1 shows that the empirical rejection rates. If the model is correct, the rejection rates are close to the nominal levels. However, if the model is contaminated by randomly replacing 10% of the data by the outlier $\mu + 3\sigma = 4$, the rejection rates increased. The rejection rates further increased by increasing the location of outliers to $\mu + 6\sigma = 7$ or increasing the contamination rates to 20%. In conclusion, the proposed goodness-of-fit test has a desirable type I error and reasonable power rates.

4. The second-order Markov model

This section develops estimation and SPC methods under the 2^{nd}-order Markov chain model, which has not been considered in the literature. The 2^{nd}-order Markov chain model is more difficult to interpret for SPC users, but it can fit well to some real dataset. Consequently, the 2^{nd}-order model is an attractive alternative to the 1^{st}-order (Markov) model, and it even provides a tool for checking a Markov property.

4.1. Model and data-generation

In this sub-section we introduce a data generation algorithm for the 2^{nd}-order Markov model.

For a sequence $\{Y_t : t = 1, \ldots, n\}$, the conditional densities under the 2^{nd}-order model is

$$g(y_t|y_{t-1}, \ldots, y_1) = g(y_t|y_{t-1}, y_{t-2}) = \frac{\partial}{\partial y_t} \Pr(Y_t \le y_t | Y_{t-1} = y_{t-1}, Y_{t-1} = y_{t-2}).$$

Hence, the probabilistic model for the sequence is specified by the joint distribution of three adjacent variables. We propose to impose a tri-variate copula function

$$F(y_t, y_{t-1}, y_{t-2}) = \Pr(Y_t \le y_t, Y_{t-1} \le y_{t-1}, Y_{t-2} \le y_{t-2}) = C_\alpha\big[G(y_t), G(y_{t-1}), G(y_{t-2})\big],$$

where $C_\alpha : [0,1]^3 \to [0,1]$ is a tri-variate copula, $\alpha \in \mathbb{R}$ is a dependence parameter, and $G(y) = \Phi\{(y-\mu)/\sigma\}$. Note that $\mu = E(Y_t)$ and $\sigma = \sqrt{\text{Var}(Y_t)}$.

For instance, the tri-variate Clayton copula is defined as

$$C_\alpha(u_1, u_2, u_3) = (u_1{}^{-\alpha} + u_2{}^{-\alpha} + u_3{}^{-\alpha} - 2)^{-1/\alpha},$$

where $\alpha > 0$ describes the correlation between Y_{t-2} and Y_{t-1}, the correlation between Y_{t-2} and Y_t, and, the correlation between Y_{t-1} and Y_t. While the model imposes a strong symmetric correlation structure, one important reason of using the tri-variate Clayton copula is its simple mathematical form allowing an explicit data-generation scheme (e.g. Rotolo, Legrand, and Van Keilegom 2013).

We develop an R function `Clayton.Markov2.DATA()` to generate data as well as an R function `Clayton.Markov2.MLE()` to compute $(\hat{\mu}, \hat{\sigma}, \hat{\alpha})$ from the data (see Section 4.2, Appendices A and B for the definitions). After installing the *Copula.Markov* package, one can easily obtain the plot by typing:

```
set.seed(1)

Y=Clayton.Markov2.DATA(n=1000,mu=0,sigma=1,alpha=8)

Clayton.Markov2.MLE(Y,plot=TRUE)
```

4.2. Maximum likelihood estimation

To derive the MLE $(\hat{\mu}, \hat{\sigma}, \hat{\alpha})$, we use the conditional densities

$$g(y_t|y_{t-1}, y_{t-2}) = \frac{C_\alpha^{[1,1,1]}\left[G(y_t), G(y_{t-1}), G(y_{t-2})\right]}{C_\alpha^{[0,1,1]}\left[1, G(y_{t-1}), G(y_{t-2})\right]} g(y_t)$$

for $t \geq 3$, and

$$g(y_2|y_1) = C_\alpha^{[0,1,1]}\left[1, G(y_2), G(y_1)\right] g(y_2)$$

where

$$C_\alpha^{[1,1,1]}(u_t, u_{t-1}, u_{t-2}) = \frac{\partial^3 C_\alpha(u_t, u_{t-1}, u_{t-2})}{\partial u_t \partial u_{t-1} \partial u_{t-2}}, C_\alpha^{[0,1,1]}(u_t, u_{t-1}, u_{t-2}) = \frac{\partial^2 C_\alpha(u_t, u_{t-1}, u_{t-2})}{\partial u_{t-1} \partial u_{t-2}}$$

The log-likelihood based on Clayton copula can be expressed as

$$\begin{aligned}
\ell(\mu, \sigma, \alpha) = {} & (n-2)\log(1+2\alpha) + \log(1+\alpha) \\
& - \left(\frac{1}{\alpha} + 3\right) \sum_{t=3}^{n} \log\left[G(y_t)^{-\alpha} + G(y_{t-1})^{-\alpha} + G(y_{t-2})^{-\alpha} - 1\right] \\
& + \left(\frac{1}{a} + 2\right) \sum_{t=4}^{n} \log\left[G(y_{t-1})^{-\alpha} + G(y_{t-2})^{-\alpha} - 1\right] \\
& - (\alpha+1) \sum_{t=1}^{n} \log G(y_t) + \sum_{t=1}^{n} \log g(y_t)
\end{aligned}$$

where $g(y_t) = \partial d G(y_t)/y_t$. Appendix C provides the derivation of $\ell(\mu, \sigma, \alpha)$, and the expressions of $C_\alpha^{[1,1,1]}(u_t, u_{t-1}, u_{t-2})$ and $C_\alpha^{[0,1,1]}(u_t, u_{t-1}, u_{t-2})$.

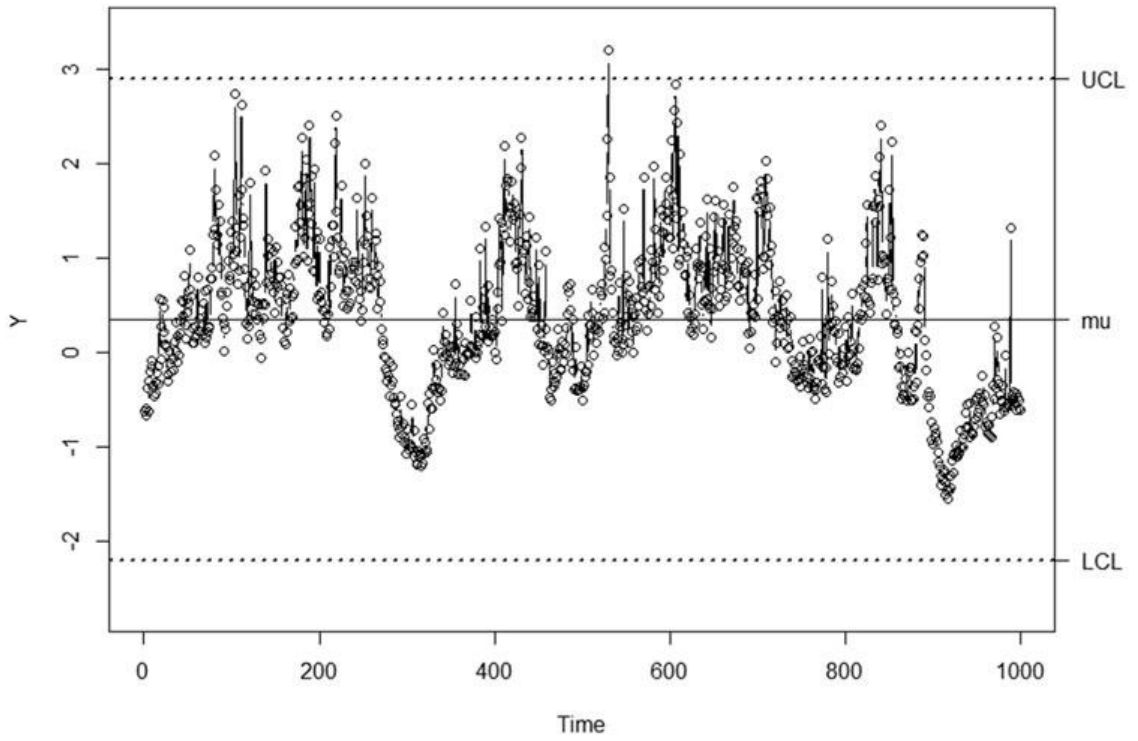

Figure 2. The plot of the 2nd-order Markov chain $\{Y_t : t = 1, \ldots, 1000\}$ under the tri-variate Clayton copula with α=8 and the marginal distribution $G \sim N(0, 1)$.

In Appendix B, we provide the definition of the R `Clayton.Markov2.MLE()` function that can numerically obtain $(\hat{\mu}, \hat{\sigma}, \hat{\alpha})$. The function applies the subroutine `nlm()` to maximize the log-likelihood with the data-driven initial values

$$\left(\bar{y} = \frac{1}{n}\sum_{t=1}^{n} y_t, \sqrt{\frac{1}{n-1}\sum_{t=1}^{n} (y_t - \bar{y})^2}, \frac{2\tau_0}{1-\tau_0} \right)$$

where

$$\tau_0 = \frac{2}{n(n-1)} \sum_{t<t*} \mathrm{sgn}\{(y_t - y_{t*})(y_{t+1} - y_{t*+1})\},$$

and where $\mathrm{sgn}(x) = -1$ for $x<0$, $\mathrm{sgn}(x) = 0$ for $x = 0$ and $\mathrm{sgn}(x) = 1$ for $x>0$. If the algorithm diverges, then it restarts after adequate randomization from $Unif(-D, D)$ in the initial values with a user specific value $D>0$. This scheme is called the randomized Newton-Raphson algorithm that has been applied to various statistical models with many parameters (Achim and Emura 2019; Emura and Pan 2017; Shih and Emura 2018; He and Emura 2019). Note that τ_0 is Kendall's tau after transforming the time series data to the paired data:

$$(y_1, y_2), (y_2, y_3), \ldots, (y_{n-1}, y_n)$$

The following example calculates the MLE:

```
> set.seed(1)
> Y=Clayton.Markov2.DATA(n=1000,mu=0,sigma=1,alpha=8)
> Clayton.Markov2.MLE(Y,plot=TRUE)
$estimates
        mu      sigma      alpha        UCL        LCL
 0.3512133  0.8471141  4.8640316  2.8925557 -2.1901291

$out_of_control
[1] 530

$gradient
[1] -4.348635e-05 -9.454880e-05 -9.170452e-06

$hessian
          [,1]       [,2]       [,3]
[1,]  754.1947  -887.7274   813.1771
[2,] -887.7274  2207.5784 -1331.2615
[3,]  813.1771 -1331.2615  1013.3597

$CM.test
[1] 0.2583273

$KS.test
[1] 0.03049542

$log_likelihood
[1] -170.0381
```

Here, we use the same data $\{Y_t : t = 1, ..., 1000\}$ as appearing in Figure 2. In the output, $estimates gives the MLE $(\hat{\mu}, \hat{\sigma}, \hat{\alpha})$, the lower control limit (LCL=$\hat{\mu}-3\hat{\sigma}$) and the upper control limit (UCL=$\hat{\mu}+3\hat{\sigma}$). Whether the MLE attains the maximum of the likelihood function or not can be confirmed by checking `$gradient` and `$hessian`. In the example, the gradients are quite close to zero, which means that the likelihood function gives a proper solution. In addition, the output shows that the Hessian matrix is negative definite since the minimum eigenvalue of the Hessian matrix is negative. This guarantees that the MLE attains the local maximum of the log-likelihood (see p. 284, Theorem 7.7.1 of Khuri 2003).

Even though n=1000 is quite large, the MLEs of $\hat{\mu}$ =0.351 and $\hat{\sigma}$ =0.847 are not close to the true values of μ =0 and σ =1. This is due to a large sampling variation caused by the strong serial correlation ($\tau = 0.8$), a reasonable phenomenon suggested

Table 2. The MSEs of estimating parameters under the 1st-order and 2nd-order Markov models. The data was generated by the Clayton copula and N($\mu = 1, \sigma = 1$).

			n =300			n =600			n =1000		
	True model	Fitted model	τ=0.2	τ=0.5	τ=0.75	τ=0.2	τ=0.5	τ=0.75	τ=0.2	τ=0.5	τ=0.75
μ	1st-order	1st-order	0.0067	0.0265	0.1319	0.0034	0.0126	0.0805	0.0020	0.0077	0.0397
		2nd-order	0.0071	0.0662	0.3896	0.0036	0.0420	0.3009	0.0022	0.0314	0.2347
σ	1st-order	1st-order	0.0027	0.0071	0.0197	0.0012	0.0030	0.0120	0.0007	0.0018	0.0068
		2nd-order	0.0027	0.0092	0.0404	0.0013	0.0047	0.0220	0.0007	0.0030	0.0150
α	1st-order	1st-order	0.0142	0.2474	6.0668	0.0064	0.1089	5.5584	0.0040	0.0671	2.6455
		2nd-order	0.5353	0.4777	21.397	0.0494	0.3548	10.165	0.0477	0.3219	3.8345
μ	2nd-order	1st-order	0.0112	0.0697	0.3192	0.0058	0.0356	0.2021	0.0035	0.0199	0.1234
		2nd-order	0.0111	0.0636	0.2466	0.0057	0.0315	0.1544	0.0034	0.0182	0.0716
σ	2nd-order	1st-order	0.0037	0.0134	0.0470	0.0018	0.0063	0.0226	0.0011	0.0037	0.0147
		2nd-order	0.0035	0.0113	0.0612	0.0017	0.0055	0.0350	0.0010	0.0030	0.0133
α	2nd-order	1st-order	0.0208	0.5146	5.4581	0.0096	0.2458	5.4641	0.0061	0.1300	5.5815
		2nd-order	0.0142	0.4871	28.042	0.0069	0.2176	20.802	0.0043	0.1172	5.5723

Table 3. The rate of choosing the model between the 1st-order and 2nd-order Markov models. The data was generated by the Clayton copula and $\mathrm{N}(\mu = 1, \sigma = 1)$.

		n =300			n =600			n =1000		
True model	Chosen model	τ=0.2	τ=0.5	τ=0.75	τ=0.2	τ=0.5	τ=0.75	τ=0.2	τ=0.5	τ=0.75
1st-order	1st-order	0.964	1.000	0.988	0.994	1.000	0.999	1.000	1.000	1.000
	2nd-order	0.036	0.000	0.012	0.006	0.000	0.001	0.000	0.000	0.000
2nd-order	1st-order	0.014	0.001	0.000	0.002	0.000	0.000	0.000	0.000	0.000
	2nd-order	0.986	0.999	1.000	0.998	1.000	1.000	1.000	1.000	1.000

The rate of choosing the 1st-order model is $\frac{1}{1000}\sum_{i=1}^{1000}\mathbf{I}(\ell_1 > \ell_2)$, and the rate of choosing the 2nd-order model is $\frac{1}{1000}\sum_{i=1}^{1000}\mathbf{I}(\ell_1 < \ell_2)$, where ℓ_k is the log-likelihood under the k-th order model.

from a plot in Figure 2. However, in the long run, the bias vanishes (see the simulation results of Section 4.4).

The function `Clayton.Markov2.MLE()` draws a control chart, including UCL=$\hat{\mu} + k\hat{\sigma}$, LCL= $\hat{\mu} - k\hat{\sigma}$ and the center line $\hat{\mu}$ (Figure 2). The default is $k = 3$ (3-sigma control limit), but the user can specify any value $k>0$. In the output, only one observation falls outside the interval [LCL, UCL]. This out-of-control signal is indeed identified from Figure 2. The value $k = 3$ means that the rate of out-of-control signals is specified at 0.27% at each time point.

4.3. Model selection

We propose a model selection method by comparing the 1st-order and 2nd-order models, and then, choosing one that fits better (higher value in the maximized log-likelihood). From the simulations below, we see that the method has a high probability to select the true model (higher log-likelihood value) if either the 1st-order or 2nd-order model is correct. In particular, we observe $\Pr(\ell_1 > \ell_2 | \text{1st} - \text{order model}) \geq 0.95$, where ℓ_k is the maximized log-likelihood under the k-th order Markov model. Even if both models are incorrect, the model with higher log-likelihood would be regarded as a better model.

4.4. Simulations

We compare the performance for the 1st–order model and 2nd–order model via a simulation study. We consider three cases: weak dependence ($\tau = 0.2$), medium dependence ($\tau = 0.5$) and strong dependence ($\tau = 0.75$) for the bivariate or tri-variate Clayton copula with marginal distribution $\mathrm{N}(\mu = 1, \sigma = 1)$ or $\mathrm{N}(\mu = 1, \sigma = 3)$. The sample sizes are set to be $n = 300$, 600 and 1000. We generated data and estimated $(\hat{\mu}, \hat{\sigma}, \hat{\alpha})$ under the 1st-order model and the 2nd-order model. We then compared their mean squared errors (MSEs) with respect to the true values. We also checked the ability of selecting the correct model.

Table 2 shows the MSEs for $(\hat{\mu}, \hat{\sigma}, \hat{\alpha})$. The MSEs are smaller under the correct model than those under the incorrect model, which are reasonable results. The difference of the MSEs between the correct and incorrect models are larger for stronger dependence. Under weak dependence, the MSEs are quite small even if the model is incorrect. The MSEs decrease as the sample size increases, meaning the consistency of the estimates. We should notice that the 2nd-order model occasionally produce an unusually large

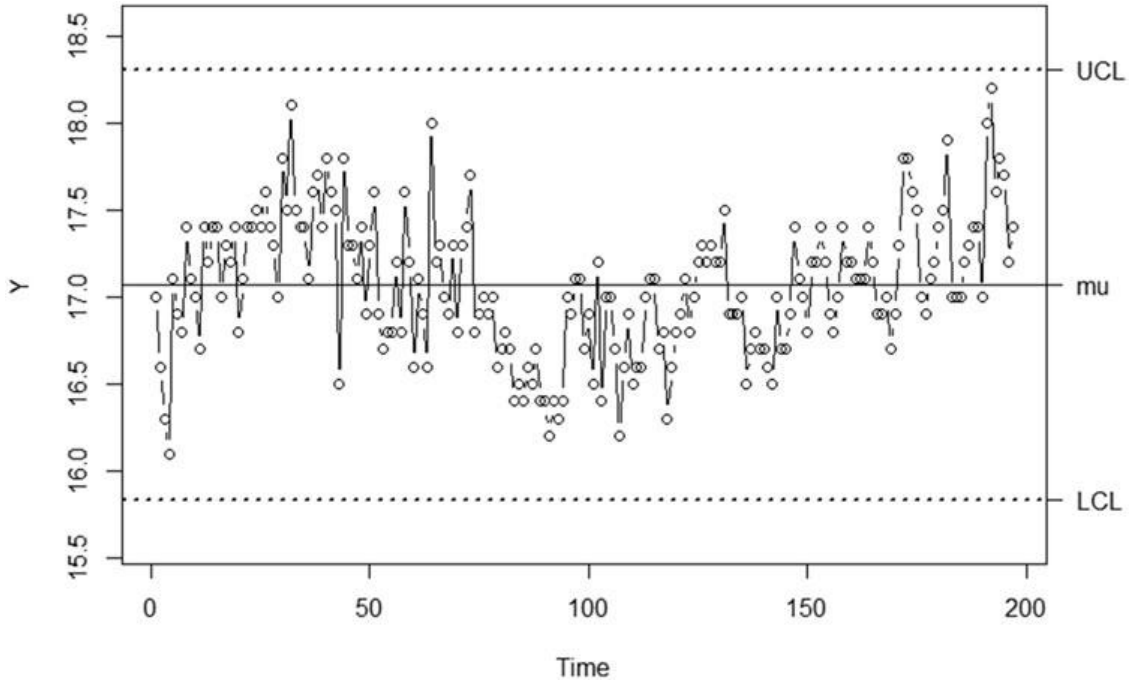

Figure 3. A control chart for chemical concentrations $\{Y_t : t = 1, \ldots, 197\}$ measured every 2 hours. The UCL and LCL are computed under the 2nd-order Markov model.

MSE for α even when the data are generated from thecorrect model. This is because strong dependence makes estimating α more difficult and unstable. Although the expectation is close to the true value, the fluctuation of the estimates produces the unusually large MSE.

Table 3 shows the performance of the proposed model selection method (Section 4.3). It reports the rate of choosing the model between the 1st-order and 2nd-order Markov models. The method has nearly 100% of selecting the correct model for n =1000. Even if the sample size is n =300, the rate of choosing the correct model is more than 95%. This results imply the model selection consistency of the proposed method.

5. Data analysis

This section analyses three datasets for illustration. The R codes for the analysis are given in Appendix D.

5.1. Chemical process data

We consider the chemical process data (Box and Jenkins 1976; Bisgaard and Kulahci 2007). The data consists of a series of chemical concentrations $\{Y_t : t = 1, \ldots, 197\}$ measured every 2 hours. Engineers use SPC to judge if the concentration level is kept within a reasonable range.

First, we applied the 1st-order Clayton Markov model (by `Clayton.Markov.MLE`), and obtained the MLE $\hat{\mu} = 17.0732223$, $\hat{\sigma} = 0.4213754$,

and $\hat{\alpha} = 1.1777489$ (Kendall's tau $\hat{\tau} = \hat{\alpha}/(\hat{\alpha}+2) = 0.37$). Control limits were LCL $= \hat{\mu} - 3\hat{\sigma} = 15.8090961$ and UCL $= \hat{\mu} + 3\hat{\sigma} = 18.3373486$.

Next, we fitted the data to the 2nd-order Clayton Markov model (by using `Clayton.Markov2.MLE`):

```
> Clayton.Markov2.MLE(Y)
$`estimate`
        mu      sigma      alpha        UCL        LCL
17.0709442  0.4123265  0.8238138 18.3079236 15.8339648

$out_of_control
[1] "NONE"

$gradient
[1] 1.453098e-07 1.649880e-07 3.694822e-09

$hessian
           [,1]        [,2]       [,3]
[1,] 406.403140   -1.151078   70.29312
[2,]  -1.151078  448.117257 -111.99199
[3,]  70.293122 -111.991989   53.19120

$CM.test
[1] 0.148302

$KS.test
[1] 0.07591838

$log_likelihood
[1] -59.32751
```

The outputs show $\hat{\mu} = 17.0709442$, $\hat{\sigma} = 0.4123265$, and $\hat{\alpha} = 0.8238138$ (Kendall's tau $\hat{\tau} = \hat{\alpha}/(\hat{\alpha}+2) = 0.29$). Control limits are LCL $= \hat{\mu} - 3\hat{\sigma} = 15.8339648$ and UCL $= \hat{\mu} + 3\hat{\sigma} = 18.3079236$.

Finally, we compared the log-likelihood for the two models: $\ell_1 = -60.07602$ (1st-order) and $\ell_2 = -59.32751$ (2nd-order). Hence, we choose the 2nd-order model for SPC. This results suggest that there may be some residual dependence that is not captured by the 1st-order model. Hence, the current chemical concentration may depend on those on previous two hours.

Figure 3 depicts a control chart drawn under the 2nd-order model. It shows that all the points are between the LCL and UCL, which implies that the process is in-control. The data clearly exhibits positive serial correlation.

5.2. Financial data

We analyze the weekly returns of S&P 500 index consisting of 500 leading companies in leading industries of the U.S. economy. Data were downloaded from FRED (Federal Reserve Economic Data) https://research.stlouisfed.org/fred2/series/SP500/downloaddata. We extracted weekly data from Jan 1st 2010 to Jan 3rd 2014 (weekly, ending Friday) and write them as $\{Y_t : t = 1, \ldots, 210\}$. The goal is to show that weekly returns stay within a reasonable range.

We first applied the 1st-order Clayton Markov model (by using `Clayton.Markov.MLE`), and obtained the MLE $\hat{\mu} = 3.28241124$, $\hat{\sigma} = 27.45415699$, and $\hat{\alpha} = 0.04422089$ (Kendall's tau $\hat{\tau} = \hat{\alpha}/(\hat{\alpha}+2) = 0.02$). Consequently, control limits were LCL$= \hat{\mu} - 3\hat{\sigma} = -79.08005974$ and, UCL$= \hat{\mu} + 3\hat{\sigma} = 85.64488222$. Since Kendall's tau

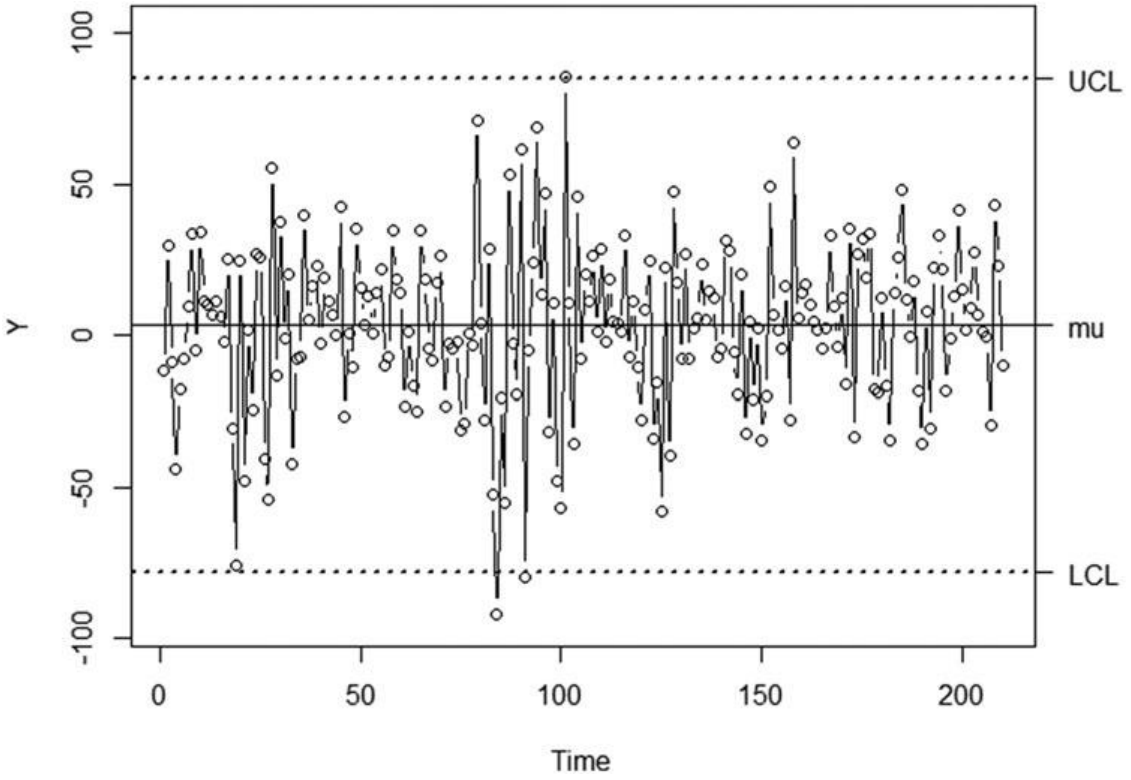

Figure 4. A control chart for weekly S&P 500 index from Jan 1st 2010 to Jan 3rd 2014 $\{Y_t : t = 1, \ldots, 210\}$. The UCL and LCL are computed under the 2nd-order Markov model.

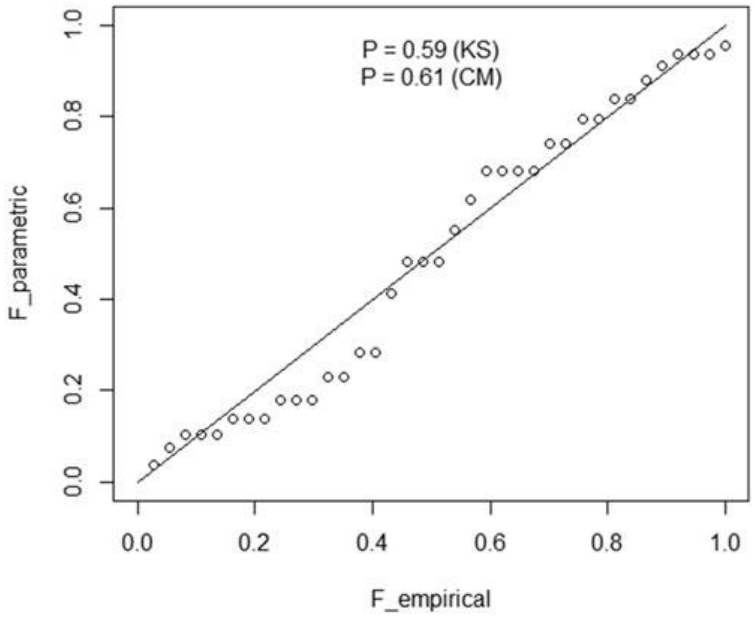

Figure 5. Goodness-of-fit test for the 1st-order Markov model under Clayton copula.

is almost zero, there is some possibility that the 1st-order Clayton Markov model cannot capture dependence structure of the data.

Next, we applied the 1st-order Joe Markov model (by using `Joe.Markov.MLE`) and obtained the MLE $\hat{\mu}$ =3.31300, $\hat{\sigma}$ =27.61220, and $\hat{\alpha}$ =2.00000 (Kendall's tau $\hat{\tau} = 1 - 4/\hat{\alpha}^2 \int_0^\infty s(1-e^{-s})^{2/\hat{\alpha}-2} e^{-2s} ds$ =0.36). Consequently, control limits are

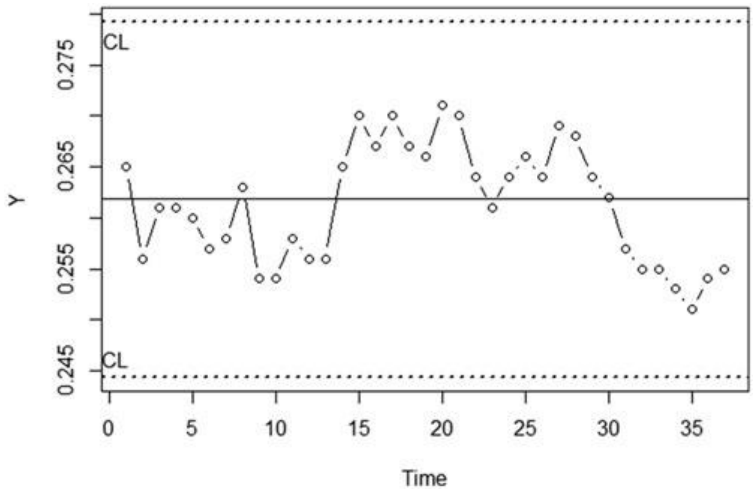

Figure 6. A control chart for BA in MLB annually returns from 1980 to 2016 $\{Y_t : t = 1, \ldots, 37\}$. The UCL and LCL are computed under the 1st-order Markov model under Clayton copula.

LCL=$\hat{\mu}-3\hat{\sigma}=-79.52359$ and, UCL=$\hat{\mu}+3\hat{\sigma}=86.14959$. It is interesting to see that Kendall's tau is now much larger than that under the Clayton copula.

Next, we fit the data to the 2nd-order model (by using `Clayton.Markov2.MLE`):

```
$`estimates`
          mu         sigma         alpha         UCL           LCL
  3.27853834  27.23464482   0.09224491  84.98247281 -78.42539612

$out_of_control
[1]  84  91 101

$gradient
[1]  5.548172e-07 -5.036709e-05  5.247140e-07

$hessian
             [,1]            [,2]             [,3]
[1,] 0.2235298   1.011451  0.1203988
[2,] 1.0114510 398.420392 -5.1577411
[3,] 0.1203988  -5.157741  3.3218271

$CM.test
[1] 0.1317453

$KS.test
[1] 0.06562425

$log_likelihood
[1] -991.992
```

The outputs show $\hat{\mu}$ =3.27853834, $\hat{\sigma}$ =27.23464482, and $\hat{\alpha}$=0.09224491 (Kendall's tau $\hat{\tau} = \hat{\alpha}/(\hat{\alpha}+2)$ =0.29). Consequently, control limits are LCL=$\hat{\mu}-3\hat{\sigma}=-78.42539612$ and UCL=$\hat{\mu}+3\hat{\sigma}$=84.98247281. Again, it is interesting to see that the see that Kendall's tau is now much larger than that under the 1st-order Clayton model.

Finally, we compared the log-likelihood for the three models: ℓ_1(Clayton)= −993.8922, ℓ_2(Clayton) = −991.992 and ℓ_1(Joe)= −1064.618. Hence, we chose the 2nd-order Clayton model for SPC.

Figure 4 depicts a control chart drawn under the 2nd-order Clayton Markov model. It shows that three points are outside the range between the LCL and UCL. Hence the process is out-of-control. The data exhibits positive but weak serial correlation.

5.3. Baseball data

A set of baseball data was analyzed as an example. The data is available on open data website: https://www.baseball-reference.com/leagues/MLB/bat.shtml. Following Kim, Baik, and Reller (2019), we extract annual records of batting average (BA) in Major League Baseball (MLB) from 1980 to 2016 and write them as $\{Y_t : t = 1, ..., 37\}$. The goal is to detect if there is a large and unusual variation of MLB statistics by fitting our method (Kim, Baik, and Reller 2019).

We first apply the 1st-order Clayton Markov model (by using `Clayton.Markov.MLE`):

```
> Clayton.Markov.MLE(Y)
$`estimates`
         mu       sigma       alpha         UCL         LCL
0.261812672 0.005793249 1.825540748 0.279192419 0.244432926

$out_of_control
[1] "NONE"

$Gradient
[1]  6.102218e-12 -4.089577e-11  1.525716e-13

$Hessian
             [,1]        [,2]        [,3]
[1,] -15875.94978   8634.3361 -35.5368485
[2,]   8634.33614 -73289.4719  84.3591025
[3,]    -35.53685     84.3591  -0.1677149

$Mineigenvalue_Hessian
[1] -74559.97

$CM.test
[1] 0.1554252

$KS.test
[1] 0.150176

$log_likelihood
[1] 153.8685
```

The outputs show the MLE $\hat{\mu} = 0.261812672$, $\hat{\sigma} = 0.005793249$, and $\hat{\alpha} = 1.825540748$ (Kendall's tau $\hat{\tau} = \hat{\alpha}/(\hat{\alpha}+2) = 0.48$). Control limits are LCL$=\hat{\mu}-3\hat{\sigma}=0.244432926$ and, UCL$=\hat{\mu}+3\hat{\sigma}=0.279192419$.

Next, we fitted the 1st-order Joe Markov model (by using `Joe.Markov.MLE`), we obtained $\hat{\mu} = 0.260683403$, $\hat{\sigma} = 0.006095821$, and $\hat{\alpha} = 2.390078566$ (Kendall's tau $\hat{\tau} = 1-4/\hat{\alpha}^2 \int_0^\infty s(1-e^{-s})^{2/\hat{\alpha}-2}e^{-2s}ds = 0.43$). Control limits are LCL$=\hat{\mu}-3\hat{\sigma}=-0.242395939$ and, UCL$=\hat{\mu}+3\hat{\sigma}=0.278970867$.

Lastly, we fitted the data to the 2nd-order Clayton Markov model (by using `Clayton.Markov2.MLE`) and obtained: $\hat{\mu} = 0.261049293$, $\hat{\sigma} = 0.005741486$, and $\hat{\alpha} = 1.368885059$ (Kendall's tau $\hat{\tau} = \hat{\alpha}/(\hat{\alpha}+2) = 0.40$). Consequently, control limits are LCL$=\hat{\mu}-3\hat{\sigma}= 0.243824833$ and UCL$=\hat{\mu}+3\hat{\sigma}= 0.278273752$.

We compared the log-likelihood for the three models: ℓ_1(Clayton)$= 153.8685$, $\ell_2 = 152.4118$ (2nd-order Clayton) and ℓ_1(Joe)$= 150.7123$. Obviously, the 1st-order Clayton Markov model is chosen for SPC.

To confirm the 1st–order Clayton Markov model as a suitable model for the dataset, we performed model diagnostic and goodness-of-fit tests. The model diagnostic plot does not give graphical evidence of rejecting the model (Figure 5). Indeed, bootstrap goodness-of-fit tests (by

`Clayton.Markov.GOF`) show little evidence for rejecting the model under the Kolmogorov-Smirnov statistics (P-value = 0.59) and Cramér-von Mises statistics (P-value = 0.61).

Figure 6 depicts a control chart drawn under the 1st-order Clayton Markov model. It shows that none of the points are outside the control limit, which implies that the process is in-control. The data exhibits positive serial correlation.

6. Conclusion and future work

In this paper, we introduce methodologies and computer algorithms to assess the goodness-of-fit of the copula-based Markov model for serially dependent time series. Proposed methods are implemented through our R package *Copula.Markov* so that users can readily perform their analysis. The package can be used to fit the data to the copula model, perform model diagnostic/goodness-of-fit analyses, select a suitable model, and perform SPC (in both 1st-order and 2nd-order models). We have demonstrated the use of the proposed methods through three data examples, including chemical data, financial data, and sports data. We believe that our new techniques in our package can greatly facilitate SPC under copula-based Markov models.

While the simulations show that the parametric bootstrap goodness-of-fit test has a desirable type I error and reasonable power rates, the asymptotic theory behind the tests remain unclear. Under independent observations, Genest and Rémillard (2008) mathematically verified the parametric bootstrap goodness-of-fit tests. Their theory cannot be applied for serially correlated observations. The derivation of the asymptotic theory often makes it possible to derive a resampling scheme based on the multiplier central limit theorem, which reduces the computational time of the usual parametric bootstrap (Emura and Konno 2012).

More complex copula models could be considered, including 3rd-order model, other copula functions, and multi-parameter copulas. To avoid unnecessary burden in practice, these complex copula models should be used only when their practical usefulness is clear in SPC. For instance, the Gaussian copula model may not be an attractive choice since the functional form of the copula is more complex than the Clayton copula and it reduces to the usual AR(1) model under the normal margin. Also, the FGM copula has a limited range of dependence, though the functional form is very simple. This is a reason why we chose the one-parameter Clayton copula model for the 1st- and 2nd-order Markov models.

This article focuses fitting the normal margins since it is the most commonly used model for SPC and is the basis of the three-sigma control limit. Nonetheless, other marginal distributions can be considered since. Sun, Lee, and Emura (2018) considers a Bayesian inference method for the t-marginal distribution under the copula-based Markov chain model motivated by heavy tailed financial data. They did not consider SPC. Under independent data, Albers and Kallenberg (2007) suggested the so-called normal power family as it is suitable for modeling and estimating control limits. Huang, Chen, and Emura (2019) proposes an np-control chart for the binomial marginal model, which requires a much more complex likelihood function than the normal marginal model even under the 1st-order Clayton copula.

Acknowledgments

We thank Professor Jong-Ming Kim for his helpful discussions on our paper when he visited our university. We also thank an anonymous reviewer for his/her comments that improved the paper.

Funding

Emura T is financially supported by Ministry of Science and Technology, Taiwan (107-2118-M-008-003-MY3).

Appendix A: R functions `clayton.Markov2.DATA`

- Description

The R function `Clayton.Markov2.DATA()` generates the datasets under a copula-based Markov chain model. The serial dependence follows the Clayton copula and the marginal (stationary) distribution follows the normal distribution.

- Usage

`Clayton.Markov2.DATA`(n,mu,sigma,alpha)

- Arguments

n: sample size
mu: mean
sigma: standard deviation (sigma $>$ 0)
alpha: association parameter (alpha $>$ 0)

- Definition

```
#####################################################################
Clayton.Markov2.DATA = function(n,mu,sigma,alpha){
  U = numeric(n)
  U[1] = runif(1, min = 0, max = 1)
  U[2]  =  ((runif(1,  min  =  0,  max  =  1)^(-alpha/(1+alpha))-1)*U[1]^(-
  alpha)+1)^(-1/alpha)
  for(i in c(3:n)){
    U[i]  =  ( runif(1,  min  =  0,  max  =  1)^(-alpha/(1+2*alpha))*(U[i-1]^(-
  alpha)+U[i-2]^(-alpha)-1)-
              U[i-1]^(-alpha)-U[i-2]^(-alpha)+2 )^(-1/alpha)
  }
  Y = qnorm(p = U, mean = mu, sd = sigma)
  return(Y)
}
```

Appendix B: R function `copula.Markov2.MLE`

- Description

The R function `Clayton.Markov2.MLE()` produces the maximum likelihood estimates and draws the Shewhart chart with k-sigma control limits (e.g., 3-sigma). The dependence model follows the Clayton copula and the marginal (stationary) distribution follows the normal distribution

- Usage

`Clayton.Markov2.MLE` (Y, k = 0.3, D = 1, plot = TRUE)

- Arguments

Y: vector of datasets
k: constant determining the length between LCL and UCL (k = 3 corresponds to 3-sigma limit)
D: diameter for U(-D, D) used in randomized Newton-Raphson
plot: boolean variable whether to plot the control chart

- Values

$estimates: Estimates of $(\hat{\mu}, \hat{\sigma}, \hat{\alpha})$, UCL and LCL.
$out_of_control: Indices for out-of-control signals.
$gradient: The gradient of the log-likelihood at the solution. They should be close to zero.
$hessian: The Hessian matrix of the log-likelihood at the solution.
$CM.test: Cramer-von Mises test statistics
$KS.test: Kolmogorov-Smirnov test statistics
$log_likelihood: The value of log-likelihood

- Definition

```
##########################################################################
Clayton.Markov2.MLE = function(Y, k = 3, D = 1, plot = TRUE, GOF=FALSE){

  n = length(Y)
  rec = NA

  ###log-likelihood
  logL = function(par){

    mu = par[1]
    sigma = exp(par[2])
    alpha = exp(par[3])

    Z = (Y - mu)/sigma
    G.yt = pnorm(q = Z, mean = 0, sd = 1)
    g.ty = 1/sigma * dnorm(x = Z, mean = 0, sd = 1)

    l = (n-2)*log(1+2*alpha) + log(1+alpha) -
      (1/alpha+3)*sum(log(G.yt[3:n]^(-alpha)+G.yt[2:(n-1)]^(-alpha)+G.yt[1:(n-
 2)]^(-alpha)-2)) +
      (1/alpha+2)*sum(log(G.yt[3:(n-1)]^(-alpha)+G.yt[2:(n-2)]^(-alpha)-1)) -
      (alpha+1)*sum(log(G.yt)) + sum(log(g.ty))

    return(-l)

  }
###initial value and randomize
tau_0 = cor(Y[2:n],Y[1:(n-1)],method="kendall")
initial = c(mean(Y),log(sd(Y)), log(ifelse(tau_0 < 0, 1, 2*tau_0/(1-tau_0))))

count = 0
repeat{
  count = count + 1
  res = try(nlm(logL, initial , hessian = TRUE))
  if( class(res)!="try-error" ){
    break;
  }else{
    initial = initial + runif(n = 3, min = -D, max = D)
  }
  if(count>100){
    return(warning("error"))
    break;
  }
}
```

```
  ###result
  mu.hat = res$estimate[1]
  sigma.hat = exp(res$estimate[2])
  alpha.hat = exp(res$estimate[3])
  UCL = mu.hat+k*sigma.hat
  LCL = mu.hat-k*sigma.hat

  ###plot
  if(plot==TRUE){
    par(mar=c(4,5,2,5))
    plot(Y~c(1:n), type = "b", ylim = c(1.1*LCL-0.1*UCL, 1.1*UCL-0.1*LCL),
         ylab = "Y", xlab = "Time", cex = 1, cex.lab = 1)

    abline(h=UCL, lty = 3, lwd = 2)
    abline(h=LCL, lty = 3, lwd = 2)
    abline(h=mu.hat, lty = 1, lwd = 1)
    axis(4, at = c(UCL, LCL, mu.hat),
         labels = c("UCL", "LCL", "mu"), cex = 1, las = 1)
  }

  ###out of control
  OC = which((Y < LCL) | (UCL < Y))
  if (length(OC) == 0) {
    OC = "NONE"
  }

  ### Goodness-of-fit ###
  F_par=pnorm( (sort(Y)-mu.hat)/sigma.hat )
  F_emp=1:n/n

  CM.test=sum( (F_emp-F_par)^2 )
  KS.test=max( abs( F_emp-F_par ) )

  if(GOF==TRUE){

  plot(F_emp,F_par,xlab="F_empirical",ylab="F_parametric",xlim=c(0,1),ylim=c(0,1))
    lines(x = c(0,1), y = c(0,1))
  }


  ##output
  MLE = c(mu.hat, sigma.hat, alpha.hat, UCL, LCL)
  names(MLE) = c("mu", "sigma", "alpha", "UCL", "LCL")

  return(list( estimates = MLE, out_of_control = OC,
               gradient = res$gradient, hessian = res$hessian,
               CM.test=CM.test,       KS.test=KS.test,log_likelihood       =       -
  logL(res$estimate)))
}

  install.packages("Copula.Markov")
  library(Copula.Markov)
```

Appendix C: Derivation of $\ell(\mu, \sigma, \alpha)$,

The likelihood function is

$$
\begin{aligned}
L(\mu,\sigma,\alpha) &= f(Y_n = y_n, \ldots, Y_1 = y_1) \\
&= f\left(Y_n = y_n | Y_{n-1} = y_{n-1}, \ldots, Y_1 = y_1\right) \times f\left(Y_{n-1} = y_{n-1} | Y_{n-2} = y_{n-2}, \ldots, Y_1 = y_1\right) \\
&\quad \times \cdots \times f\left(Y_2 = y_2 | Y_1 = y_1\right) \times f(Y_1 = y_1) \\
&= \prod_{t=3}^{n} f\left(Y_t = y_t | Y_{t-1} = y_{t-1}, Y_{t-2} = y_{t-2}\right) \times f\left(Y_2 = y_2 | Y_1 = y_1\right) \times f(Y_1 = y_1) \\
&= \prod_{t=3}^{n} \frac{C_\alpha^{[1,1,1]}\left[G(y_t), G(y_{t-1}), G(y_{t-2})\right]}{C_\alpha^{[0,1,1]}\left[1, G(y_{t-1}), G(y_{t-2})\right]} \times C_\alpha^{[0,1,1]}\left[1, G(y_2), G(y_1)\right] \times \prod_{t=1}^{n} g(y_t)
\end{aligned}
$$

then the log-likelihood function is

$$\ell(\mu,\sigma,\alpha)=\sum_{t=3}^{n}\log C_{\alpha}^{[1,1,1]}\left[G(y_t),G(y_{t-1}),G(y_{t-2})\right]-\sum_{t=3}^{n}\log C_{\alpha}^{[0,1,1]}\left[1,G(y_{t-1}),G(y_{t-2})\right]$$
$$+\log C_{\alpha}^{[0,1,1]}\left[1,G(y_2),G(y_1)\right]+\sum_{t=1}^{n}\log g(y_t)$$

where the expressions of $C_{\alpha}^{[1,1,1]}(u_t,u_{t-1},u_{t-2})$ and $C_{\alpha}^{[0,1,1]}(u_t,u_{t-1},u_{t-2})$ under are

$$\begin{aligned}C_{\alpha}^{[1,1,1]}(u_t,u_{t-1},u_{t-2})&=\frac{\partial^3}{\partial u_t\partial u_{t-1}\partial u_{t-2}}C_{\alpha}(u_t,u_{t-1},u_{t-2})\\&=(1+\alpha)(1+2\alpha)(u_t^{-\alpha}+u_{t-1}^{-\alpha}+u_{t-2}^{-\alpha}-2)^{-1/\alpha-3}u_t^{-\alpha-1}u_{t-1}^{-\alpha-1}u_{t-2}^{-\alpha-1}\end{aligned}$$

and

$$\begin{aligned}C_{\alpha}^{[0,1,1]}(u_t,u_{t-1},u_{t-2})&=\frac{\partial^2}{\partial u_{t-1}\partial u_{t-2}}C_{\alpha}(u_t,u_{t-1},u_{t-2})\\&=(1+\alpha)(u_t^{-\alpha}+u_{t-1}^{-\alpha}+u_{t-2}^{-\alpha}-2)^{-1/\alpha-2}u_{t-1}^{-\alpha-1}u_{t-2}^{-\alpha-1}\end{aligned}.$$

Thus the log-likelihood function is

$$\ell(\mu,\sigma,\alpha)=(n-2)\log(1+2\alpha)+\log(1+\alpha)-\left(\frac{1}{\alpha}+3\right)\sum_{t=3}^{n}\log\left[G(y_t)^{-\alpha}+G(y_{t-1})^{-\alpha}+G(y_{t-2})^{-\alpha}-1\right]$$
$$+\left(\frac{1}{a}+2\right)\sum_{t=4}^{n}\log\left[G(y_{t-1})^{-\alpha}+G(y_{t-2})^{-\alpha}-1\right]-(\alpha+1)\sum_{t=1}^{n}\log G(y_t)+\sum_{t=1}^{n}\log g(y_t)$$

Appendix D: R codes for the data analyses

```
##
set.seed(1)
Y=Clayton.Markov2.DATA(n=1000,mu=0,sigma=1,alpha=8)
Clayton.Markov2.MLE(Y,plot=TRUE)

#Chemical
Y=c(17.0, 16.6, 16.3, 16.1, 17.1, 16.9, 16.8, 17.4, 17.1, 17.0, 16.7,
    17.4, 17.2, 17.4, 17.4, 17.0, 17.3, 17.2, 17.4, 16.8, 17.1, 17.4,
    17.4, 17.5, 17.4, 17.6, 17.4, 17.3, 17.0, 17.8, 17.5, 18.1, 17.5,
    17.4, 17.4, 17.1, 17.6, 17.7, 17.4, 17.8, 17.6, 17.5, 16.5, 17.8,
    17.3, 17.3, 17.1, 17.4, 16.9, 17.3, 17.6, 16.9, 16.7, 16.8, 16.8,
    17.2, 16.8, 17.6, 17.2, 16.6, 17.1, 16.9, 16.6, 18.0, 17.2, 17.3,
    17.0, 16.9, 17.3, 16.8, 17.3, 17.4, 17.7, 16.8, 16.9, 17.0, 16.9,
    17.0, 16.6, 16.7, 16.8, 16.7, 16.4, 16.5, 16.4, 16.6, 16.5, 16.7,
    16.4, 16.4, 16.2, 16.4, 16.3, 16.4, 17.0, 16.9, 17.1, 17.1, 16.7,
    16.9, 16.5, 17.2, 16.4, 17.0, 17.0, 16.7, 16.2, 16.6, 16.9, 16.5,
    16.6, 16.6, 17.0, 17.1, 17.1, 16.7, 16.8, 16.3, 16.6, 16.8, 16.9,
    17.1, 16.8, 17.0, 17.2, 17.3, 17.2, 17.3, 17.2, 17.2, 17.5, 16.9,
    16.9, 16.9, 17.0, 16.5, 16.7, 16.8, 16.7, 16.7, 16.6, 16.5, 17.0,
    16.7, 16.7, 16.9, 17.4, 17.1, 17.0, 16.8, 17.2, 17.2, 17.4, 17.2,
    16.9, 16.8, 17.0, 17.4, 17.2, 17.2, 17.1, 17.1, 17.1, 17.4, 17.2,
    16.9, 16.9, 17.0, 16.7, 16.9, 17.3, 17.8, 17.8, 17.6, 17.5, 17.0,
    16.9, 17.1, 17.2, 17.4, 17.5, 17.9, 17.0, 17.0, 17.0, 17.2, 17.3,
    17.4, 17.4, 17.0, 18.0, 18.2, 17.6, 17.8, 17.7, 17.2, 17.4)

  #Finacial
  Y=c(-11.38, 29.88, -8.95, -44.27,-17.89, -7.68, 9.32,  33.66, -4.68, 34.21, 11.29,
    9.91,  6.69,  11.51,  6.27,  -2.24,  25.15,  -30.59,  -75.81,  24.80,  -47.99,  1.72,  -
    24.53,  26.72,  25.91,  -40.75,  -54.18,  55.38,  -13.08,  37.78,  -1.06,  20.04,  -42.39,  -
    7.56,  -7.10,  39.92,  5.04,  16.04,  23.08,  -2.43,  18.91,  11.04,  6.89,  0.18,  42.59,  -
    26.64,  0.52,  -10.33,  35.31,  15.69,  3.51,  12.86,  0.87,  13.86,  21.74,  -9.89,  -7.01,
```

```
  34.53, 18.28, 13.86, -23.13, 1.27, -16.87, -25.08, 34.60, 18.61, -4.24, -8.49,
  17.70, 26.23, -23.41, -2.43, -4.50, -2.17, -30.94, -29.18, 0.52, -3.05, 71.22,
  4.13, -27.66, 28.88, -52.74, -92.09, -20.57, -55.28, 53.25, -2.83, -19.74, 61.78,
  -79.58, -5.01, 24.04, 69.12, 13.67, 46.84, -31.86, 10.62, -48.20, -56.98, 85.61,
  10.91, -35.53, 45.67, -7.73, 20.21, 11.28, 26.29, 0.95, 28.57, -2.26, 18.59, 4.51,
  3.89, 1.24, 33.30, -7.06, 11.36, -10.39, -27.82, 8.27, 24.83, -34.26, -15.71, -
  58.17, 22.60, -39.78, 47.62, 17.18, -7.82, 27.14, -7.48, 2.10, 5.88, 23.31, 5.02,
  14.88, 12.29, -7.03, -4.55, 31.34, 27.85, -5.62, -19.48, 20.26, -32.34, 4.60, -
  21.25, 2.26, -34.35, -19.97, 49.27, 7.03, 1.89, -4.49, 16.57, -27.72, 64.04, 5.58,
  13.93, 16.98, 10.21, 4.76, 1.86, -4.19, 2.60, 32.98, 9.52, -3.81, 12.30, -15.91,
  35.57, -33.60, 26.99, 32.18, 19.28, 33.77, -17.87, -18.86, 12.64, -16.65, -34.30,
  13.85, 25.61, 48.30, 11.90, -0.44, 18.02, -18.2, -35.59, 7.67, -30.53, 22.20,
  32.82, 21.92, -18.16, -1.25, 12.70, 41.30, 15.27, 1.87, 8.97, 27.57, 6.58, 1.05, -
  0.72, -29.77, 43.00, 23.08, -10.03)

#BA data
Y = c(0.265, 0.256, 0.261, 0.261, 0.260, 0.257, 0.258, 0.263, 0.254, 0.254,
      0.258, 0.256, 0.256, 0.265, 0.270, 0.267, 0.270, 0.267, 0.266, 0.271,
      0.270, 0.264, 0.261, 0.264, 0.266, 0.264, 0.269, 0.268, 0.264, 0.262,
      0.257, 0.255, 0.255, 0.253, 0.251, 0.254, 0.255)

Clayton.Markov.MLE(Y=Y)
Clayton.Markov2.MLE(Y=Y)
Joe.Markov.MLE(Y=Y)

joe.int = function(x){

  alpha = 2.39

  return(
    x*(1-exp(-x))^(2/alpha-2)*exp(-2*x)/alpha^2
  )

}
intres = integrate(joe.int,0,Inf)
1-4*intres$value

Clayton.Markov.GOF(Y = Y)
```

References

Abegaz, F., and U. V. Naik-Nimbalkar. 2008. Dynamic copula-based Markov time series. *Communications in Statistics–Theory and Method* 37 (15):2447–60. doi:10.1080/03610920801931846.

Achim, D., and T. Emura. 2019. *Analysis of doubly truncated data, an introduction, JSS research series in statistics*. Singapore: Springer.

Albers, W., and W. C. Kallenberg. 2007. Shewhart control charts in new perspective. *Sequential Analysis* 26 (2):123–51. doi:10.1080/07474940701246992.

Bisgaard, S., and M. Kulahci. 2007. Quality quandaries, practical time series modeling II. *Quality Engineering* 19 (4):393–400. doi:10.1080/08982110701456560.

Box, G. E. P., and G. M. Jenkins. 1976. *Time series analysis: Forecasting and control*. 2nd ed. Englewood Cliffs, NJ: Prentice Hall.

Box, G., and S. Narasimhan. 2010. Rethinking statistics for quality control. *Quality Engineering* 22 (2):60–72. doi:10.1080/08982110903510297.

Chen, X., and Y. Fan. 2006. Estimation of copula-based semiparametric time series models. *Journal of Econometrics* 130 (2):307–35. doi:10.1016/j.jeconom.2005.03.004.

Darsow, W. F., B. Nguyen, and E. T. Olsen. 1992. Copulas and Markov Processes. *Illinois Journal of Mathematics* 36 (4):600–42. doi:10.1215/ijm/1255987328.

Domma, F., S. Giordano, and P. F. Perri. 2009. Statistical modeling of temporal dependence in financial data via a copula function. *Communications in Statistics - Simulation and Computation* 38 (4):703–28. doi:10.1080/03610910802645321.

Emura, T., and Y. H. Chen. 2016. Gene selection for survival data under dependent censoring, a copula-based approach. *Statistical Methods in Medical Research* 25 (6):2840–57.

Emura, T., and Y. Konno. 2012. A goodness-of-fit tests for parametric models based on dependently truncated data. *Computational Statistics & Data Analysis* 56 (7):2237–50.
Emura, T., T. H. Long, and L. H. Sun. 2017. R routines for performing estimation and statistical process control under copula-based time series models. *Communications in Statistics - Simulation and Computation* 46 (4):3067–87.
Emura, T., S. Matsui, and H. Y. Chen. 2019. compound.Cox: Univariate feature selection and compound covariate for predicting survival. *Computer Methods and Programs in Biomedicine* 168:21–37. doi:10.1016/j.cmpb.2018.10.020.
Emura, T., S. Matsui, and V. Rondeau. 2019. *Survival analysis with correlated endpoints, joint Frailty-Copula models, JSS research series in statistics*. Singapore: Springer.
Emura, T., M. Nakatochi, S. Matsui, H. Michimae, V. Rondeau. 2018. Personalized dynamic prediction of death according to tumour progression and high-dimensional genetic factors: meta-analysis with a joint model. *Statistical Methods in Medical Research* 27 (9):2842–58.
Emura, T., M. Nakatochi, K. Murotani, and V. Rondeau. 2017. A joint frailty-copula model between tumour progression and death for meta-analysis. *Statistical Methods in Medical Research* 26 (6):2649–66.
Emura, T., and C. H. Pan. 2017. Parametric maximum likelihood inference and goodness-of-fit tests for dependently left-truncated data, a copula-based approach. *Statistical Papers* doi: 10.1007/s00362-017-0947-z.
Erkal Sonmez, O., and A. Baray. 2019. On Copula Based Serial Dependence in Statistical Process Control. In *Industrial engineering in the big data era. Lecture notes in management and industrial engineering*. eds. Calisir F., Cevikcan E., Camgoz Akdag H. Cham: Springer.
Genest, C., and B. Rémillard. 2008. Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models. Annales de L'institut Henri Poincaré, Probabilités et Statistiques 44 (6):1096–127. doi:10.1214/07-AIHP148.
He, Z., and T. Emura. 2019. Likelihood inference under the COM-Poisson cure model for survival data - computational aspects. *Journal of Chinese Statistical Association* 57:1–42.
Huang, X. W., W. R. Chen, and T. Emura. 2019. A control chart using a copula-based Markov chain for attribute data, revision under. *Econometrics and Statistics*
Joe, H. 1993. Parametric families of multivariate distributions with given margins. *Journal of Multivariate Analysis* 46 (2):262–82. doi:10.1006/jmva.1993.1061.
Joe, H. 1997. *Multivariate models and dependence concepts*. Boca Raton, FL: Chapman & Hall/CRC.
Knoth, S., and W. Schmid. 2004. Control charts for time series: A review. In *Frontiers in statistical quality control 7*. ed. Lenz, H.-J. Verlag Berlin Heidelberg: Springer.
Khuri, A. I. 2003. *Advanced calculus with applications in statistics*. 2nd ed., New York: Wiley.
Kim, J. M., and J. Baik. 2018. Anomaly detection in sensor data. *Journal of Applied Reliability* 18 (1):20–32.
Kim, J. M., J. Baik, and M. Reller. 2018. Detecting the change of variance by using conditional distribution with diverse copula functions. In *Proceedings of the pacific rim statistical conference for production engineering*. (pp. 145–154). Singapore: Springer.
Kim, J. M., J. Baik, and M. Reller. 2019. Control charts of mean and variance using copula Markov SPC and conditional distribution by copula. *Communications in Statistics: Simulation and Computation* :1. doi:10.1080/03610918.2018.1547404.
Long, T. H., and T. Emura. 2014. A control chart using copula-based Markov chain models. *Journal of the Chinese Statistical Association* 52 (4): 466–96.
Montgomery, D. C. 2009. *Statistical quality control*. 6th ed. New York: Wiley.
Nelsen, R. B. 2006. *An introduction to copulas*. 2nd ed. Verlag, New York: Springer Series in Statistics, Springer.
Rotolo, F., C. Legrand, and I. Van Keilegom. 2013. A simulation procedure based on copulas to generate clustered multi-state survival data. Computer Methods and Programs in Biomedicine 109 (3):305–12. doi:10.1016/j.cmpb.2012.09.003.
Rotolo, F., X. Paoletti, and S. Michiels. 2018. surrosurv: An R package for the evaluation of failure time surrogate endpoints in individual patient data meta-analyses of randomized clinical

trials. *Computer Methods and Programs in Biomedicine* 155:189–98. doi:10.1016/j.cmpb.2017.12.005.

Shih, J. H., and T. Emura. 2018. Likelihood-based inference for bivariate latent failure time models with competing risks under the generalized FGM copula. *Computational Statistics* 33 (3): 1293–23.

Sun, L. H., C. S. Lee, and T. Emura. 2018. A Bayesian inference for time series via copula-based Markov chain models. *Communications in Statistics*:1. doi:10.1080/03610918.2018.1529241.

Wieringa, J. E. 1999. *Statistical process control for serially correlated data*. PhD thesis. University of Groningen.