Jurnal Teknologi

Full Paper

PARAMETRIC ESTIMATION METHODS FOR BIVARIATE COPULA IN RAINFALL APPLICATION

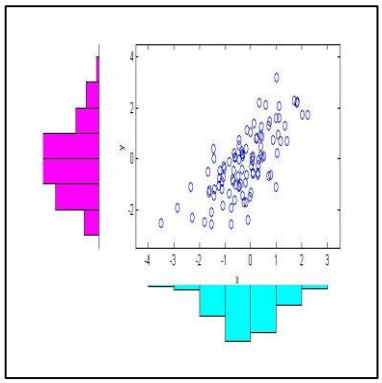
Rahmah Mohd Lokoman, Fadhilah Yusof*

Mathematical Department, Faculty of Science, Universiti Teknologi Malaysia, 81300 UTM Johor Bahru, Johor, Malaysia

Article history
Received
16 November 2017
Received in revised form
19 July 2018
Accepted
1 August 2018
Published online
15 December 2018

*Corresponding author
fadhilahy@utm.my

Graphical abstract



Abstract

This study focuses on the parametric methods: maximum likelihood (ML), inference function of margins (IFM), and adaptive maximization by parts (AMBP) in estimating copula dependence parameter. Their performance is compared through simulation and empirical studies. For the empirical study, 44 years of daily rainfall data of Station Kuala Krai and Station Ulu Sekor were used. The correlation of the two stations is statistically significant at 0.4137. The results from the simulation study show that when the sample size is small (n <1000) for correlation level less than 0.80, IFM has the best performance. While, when the sample size is large (n ≥ 1000) for any correlation level, AMBP has the best performance. The results from the empirical study also show that AMBP has the best performance when the sample size is large. Thus, in order to estimate a precise Copula dependence parameter, it can be concluded that for parametric approaches, IFM is preferred for small sample size and has correlation level less than 0.80 and AMBP is preferred for larger sample size and for any correlation level. The results obtained in this study highlight the importance of estimating the dependence structure of the hydrological data. By using the fitted copula, the Malaysian Meteorological Department will be able to generate hydrological events for a system performance analysis such as flood and drought control system.

Keywords: Bivariate copula, maximum likelihood, Inference function of margins, adaptive maximization by parts, rainfall

Abstrak

Kajian ini memberi tumpuan kepada kaedah parametrik: kebolehjadian maksimum (ML), fungsi taksiran marginal (IFM), dan penyesuaian pengoptimuman bahagian demi bahagian (AMBP) dalam menganggarkan parameter bersandar Copula. Prestasi mereka telah dibandingkan melalui kajian simulasi dan kajian empirikal. Untuk kajian empirikal, data hujan harian selama 44 tahun di Stesen Kuala Krai dan Stesen Ulu Sekor digunakan. Hubungan kedua-dua stesen adalah signifikan secara statistik pada 0.4137. Hasil daripada kajian simulasi menunjukkan bahawa apabila saiz sampel kecil (n <1000) untuk tahap korelasi kurang dari 0.80, IFM mempunyai prestasi terbaik. Manakala, apabila saiz sampelnya besar (n ≥ 1000) untuk mana-mana tahap korelasi, AMBP mempunyai prestasi terbaik. Hasil daripada kajian empirikal juga menunjukkan bahawa AMBP mempunyai prestasi terbaik apabila saiz sampelnya besar. Oleh itu, untuk menganggarkan parameter bersandar Copula yang tepat, dapat disimpulkan bahawa untuk pendekatan parametrik, IFM adalah kaedah yang bagus untuk saiz sampel yang kecil dan mempunyao korelasi kurang dari 0.80 dan AMBP untuk saiz sampel yang lebih besar untuk mana-mana tahap korelasi. Keputusan yang diperolehi dalam kajian ini menunjukkan pentingnya menganggar struktur ketersandaran data hidrologi. Dengan menggunakan taburan Copula yang terpilih, Jabatan Meteorologi Malaysia boleh

menjana peristiwa hidrologi untuk membuat analisis prestasi sistem seperti sistem kawalan banjir dan kemarau.

Kata kunci: Copula bivariat, Kebolehjadian Maksimum, Fungsi Taksiran Marginal, penyesuaian pengoptimuman bahagian demi bahagian, hujan

1.0 INTRODUCTION

Copula method was introduced by Sklar [1]. A copula function is a joint distribution function of a combination of two or more uniform marginal distributions. This method can overcome the limitations of the traditional approach because it allows us to specify any distribution function to the marginal distributions and then choose any copula to construct the dependence structure of the variables. In the work of Zhang and Singh [2], they have proved that the copula method is able to derive bivariate joint distributions of rainfall variables that have different marginal distributions and without assuming the variables to be normal or independent. Yee *et al*. [3] stated that many different copula families that are able to cover a wide scope of dependence structures have been proposed and developed, for example, Archimedean, Gaussian, and Student's t copula families.

To determine a specified copula structure that is fitted with the marginal variables, the parameter of the copula function need to be estimated first. There are many parameter estimation methods that have been proposed and developed for estimating the dependence parameter of the copula. These methods are classified into three categories, parametric approaches, semiparametric approaches, and nonparametric approaches. Some comparison studies such as Kim *et al*. [4], Kim *et al*. [5], Kojadinovic and Yan [6], Lawless *et al*. [7] and Nagler *et al*. [8] were done to compare the performance of all the copula parameter estimations methods. A study by Taheri *et al*. [9] has applied parametric, semiparametric and nonparametric methods, for estimating the dependence parameter Θ and other related parameters for bivariate situations in presence of outliers.

In parametric approaches, the marginal distributions are assumed to follow a parametric distribution. The parameters of interest are marginal parameters and copula dependence parameter. Parametric methods are popular because they estimate the estimator precisely. Parametric approaches allow the estimation process assuming parametric distributions for both the copula and the marginal. There are three estimation methods that have been reviewed under this approach. The methods are maximum likelihood (ML) estimation, inference function of margins (IFM) and maximization by parts (MBP).

Maximum likelihood (ML) estimation is a direct maximization method to estimate the marginal and copula parameters simultaneously. This direct maximization method is a common method to estimate the copula estimator. The ML estimator is also the most efficient estimator for the copula dependence parameter, θ. This is because this method is asymptotically normal and consistent under the common regularity conditions. However, in real application, it is difficult to maximize the log-likelihood function with respect to the marginal parameters, α, β and dependence parameter, θ simultaneously. Therefore, a numerical iterative method such as Newton-Raphson is used to find the ML estimator. For the bivariate copula function which has a simple dependence structure, ML estimation is possible to be applied. However, when there is a high dimensional parameter, the optimization algorithm for the iterative method becomes computationally difficult and intensive. Dupuis [10] and Zhang *et al*. [11] agree with this limitation of the ML estimation.

Joe and Xu [12] suggested an estimation method called inference function of margins (IFM). The estimation of the marginal parameters and copula parameter is done separately by this IFM method. It is also asymptotically normal and consistent under the common regularity conditions. Thus, it makes the IFM estimator, $\hat{\theta}_{IFM}$ efficient similarly similarly to the ML estimator, $\hat{\theta}_{MLE}$. Joe [13] said that this IFM method makes a huge contribution to the computational practicality since this estimation method can be applied when the ML estimation method is computationally too difficult. They also said that the main purpose of the proposed IFM method is only for the computational implementation, not for the theoretical analysis.

The main advantage of this IFM method is it is computationally efficient than ML estimation because it does not estimate the marginal and dependence parameters simultaneously. The estimator of the IFM method is efficient if the bivariate random variables have no dependency or the dependency level is low. IFM estimator can be efficient similarly to the ML estimator because both methods estimate both marginal and copula parameters. However, according to Zhang *et al*. [11] and Song *et al*. [14], IFM estimator can be less efficient compared to ML estimator because IFM estimates marginal and

copula parameters separately. Meanwhile ML estimates marginal and copula parameters simultaneously. The first step in the IFM method only considers marginal parameters but disregards the dependence level that may exist between the marginal random variables.

To overcome the loss of the copula estimator efficiency in the first step of IFM method, Song *et al*. [14] recommended and examined a simple new algorithm that maximizes the full log-likelihood function of copula by parts iteratively. This method is called as maximization by parts (MBP). This new algorithm iteratively solves a score equation to estimate the parameters. Song *et al*. [14] decomposed the full log-likelihood equation of copula into two parts or two models. Where the first part is called the working model in which the model consists of only the marginal parameters. While the second part consists of both marginal and copula parameters and this part is called the error model. Consequently, the decomposition makes the marginal log-likelihood model as the working and the copula log-likelihood model as the error model. The iterative MBP algorithm proposed by Song *et al*. [14] is based on bivariate Gaussian copula.

Silvennoinen and Teräsvirta [15] also said that MBP method reduces the computational problem as instead of maximizing the whole log-likelihood at once, MBP method divides the maximization problem into parts. However, Frazier and Renault [16] said that the limitation of this method is that it is too time-consuming and will be difficult when the variables have high correlation values and larger sample sizes. To overcome the limitation of MBP method by Song *et al*. [14], Zhang *et al*. [11] proposed an adaptive maximization by parts (AMBP) algorithm based on Meta *t* distributions to improve the MBP method by Song *et al*. [14].

Though, in the hydrological analysis, Kendall's tau method which is classified under semiparametric approach is the most popular method that have been used for estimating the bivariate copula parameter as can be seen in studies by Zhang and Singh [2], Ariff *et al*. [17], Requena *et al*. [18] and Yusof *et al*. [19]. This is because it has a closed form of one-to-one relationship between rank correlation, tau (τ) and the copula parameter, θ which has made the estimation process become easier. Vandenberghe *et al*. [20] and Chen *et al*. [21] also preferred to use Kendall's tau method than ML estimation because it is easier to estimate the copula parameter based on Kendall's tau rank correlation coefficient rather than finding the fitted marginal distributions and maximizing a log-likelihood function that leads to a complicated algorithm. Still, parametric approaches estimates are more precise than semi-parametric approaches. This is because the parametric approaches consider the marginal parameters but semiparametric approaches ignore the marginal parameters. According to Kim *et al*. [4], Kojadinovic and Yan [6], a precise copula parameter can be estimated if the marginal parameters are considered. In addition, the most common parametric approach used in the hydrological analysis are Maximum likelihood (ML) estimation and Inference Function of Margins (IFM). However, studies that implement adaptive maximization by parts (AMBP) are atypical to find in hydrologic application literature.

Therefore, this study focuses on the application of parametric approaches: maximum likelihood (ML) estimation, inference function of margins (IFM) and adaptive maximization by parts (AMBP) in estimating the copula dependence parameter. The estimation performance of the three parametric estimation methods is compared in the simulation and empirical studies. This paper is organized as follows. Section 2 gives the scope of the study. Section 3 describes the methodology and the procedures involve in the simulation and empirical studies. Section 4 presents and explains the results of the simulation and empirical studies. Lastly, Section 5 gives the conclusions of this study.

2.0 METHODOLOGY

2.1 Scopes of the Study

In the simulation study, simulation data were generated from Clayton copula [22] as the true copula with four different values of true copula parameter dependence that corresponds to the Kendall's tau values at τ = 0.20, 0.50, 0.60, and 0.80. The sample sizes of the generated data are set at n = 50, 100, 1000, and 5000. 500 repetitions of data generation and estimation process are done for each combination of different sample size, n and copula dependence level, θ.

While, for the empirical study, rainfall data are used as the empirical data. The rainfall data are selected from two Kelantan rain gauge stations which are located in the north-east of Peninsular Malaysia. The selected rain gauge stations are Station Kuala Krai, 5522047 (Station A) and Station Ulu Sekor, 5520001 (Station B). The location of these two stations are shown in Figure 1. Forty-four years (1970-2014) of daily rainfall data from both stations were obtained from the Malaysian Meteorological Department and Drainage and Irrigation Department.

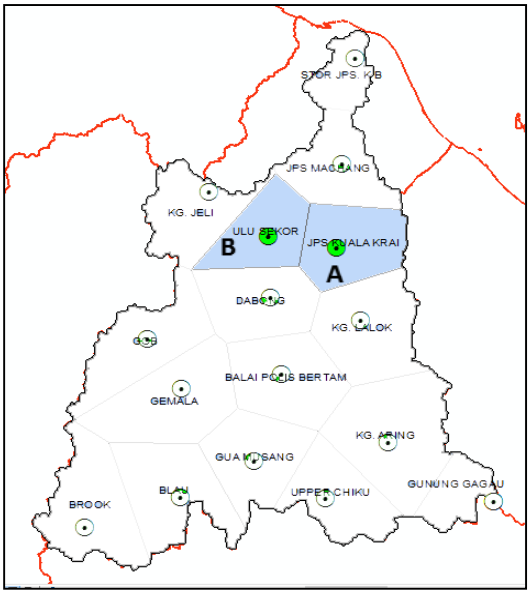


Figure 1 The location of Station A and B with their respective neighbouring rain gauge stations

2.2 Copula

Let two variables *X* and *Y* be the random variables that have marginal cumulative distribution function (CDF), $F_X(x;\alpha)$ and $F_Y(y;\beta)$ respectively with α and β as the marginal parameters for random variables *X* and *Y* respectively. Then, the joint CDF of random variables *X* and *Y* can be expressed in copula function as follows:

$$\begin{aligned} C(u,v;\ \alpha,\beta,\theta) &= C_\theta[F_X(x\,;\alpha)\,,F_Y(y\,;\beta);\theta] \\ &= H(x,y\,;\ \alpha,\beta,\theta) \end{aligned} \tag{1}$$

where v and u are the CDF of *Y* and *X* respectively.

Consequently, the joint probability density function of copula is

$$\begin{aligned} &h(x,y;\alpha,\beta,\theta) \\ &= \frac{\partial^2}{\partial x \partial y} H(x,y\,;\ \alpha,\beta,\theta) \\ &= c[F_X(x\,;\alpha)\,,F_Y(y\,;\beta);\theta]\cdot\ f_X(x\,;\alpha)\ \cdot\ f_Y(y\,;\beta) \end{aligned} \tag{2}$$

where

$$\begin{aligned} c[F_X(x\,;\alpha)\,,F_Y(y\,;\beta);\theta] &= c(u,v;\theta) \\ &= \frac{\partial^2}{\partial u \partial v} C(u,v;\theta) \end{aligned} \tag{3}$$

is the PDF of the copula function and $f_X(x\,;\alpha)$ and $f_Y(y\,;\beta)$ are the PDF of random variables of *X* and *Y* respectively.

Further, the detailed steps to get the estimator of dependence parameter, θ by maximum likelihood (ML) estimation, inference function of margins (IFM) and adaptive maximization by parts (AMBP) are explained in the following sections.

2.3 Maximum Likelihood Estimation

Maximum likelihood (ML) estimation is a direct maximization method to estimate the marginal and copula parameters simultaneously by maximizing the log-likelihood of the copula joint PDF

The steps involved in the ML estimation are described as follows:

Step 1: Find the likelihood function of equation (2). The likelihood form of the likelihood function with the random variables $\{(x_i)\}_{i=1}^n$ and $\{(y_i)\}_{i=1}^n$is written as follows:

$$\begin{aligned} &L\ (\alpha,\beta,\theta) \\ &= \prod_{i=1}^{n} c[F_X(x_i;\alpha),F_Y(y_i;\beta);\theta]\ \cdot f_X(x_i;\alpha)\cdot f_Y(y_i;\beta) \end{aligned} \tag{4}$$

Step 2: Find the log-likelihood function of equation (4). The log-likelihood form is

$$\begin{aligned} &\ln L\ (\alpha,\beta,\theta) \\ &= \sum_{i=1}^{n} \ln c[F_X(x_i;\alpha),F_Y(y_i;\beta);\theta] + [\sum_{i=1}^{n} \ln f_X(x_i;\alpha) \\ &\quad + \sum_{i=1}^{n} \ln f_Y(y_i;\beta)] \end{aligned} \tag{5}$$

Step 3: Maximize the full copula log-likelihood function (Eq. 5) with an expression as below

$$\hat{\alpha},\hat{\beta},\hat{\theta} = \ \text{argmax}\ (\ln L\ (\alpha,\beta,\theta)) \tag{6}$$

To maximize the log-likelihood function (Eq. 5), this study used optimization algorithm (the iterative method) since it is difficult and complicated to solve the nonlinear simultaneous equation manually.

2.4 Inference Function of Margins

To implement this method, the log-likelihood function (Eq. 5) is separated into two parts, a marginal and copula log-likelihood model. The log-likelihood function (Eq. 5) is also written as

$$\ell(\alpha,\beta,\theta) = \ \ell_m(\alpha,\beta)\ +\ \ell_c(\alpha,\beta,\theta) \tag{7}$$

where

$$\ell_m(\alpha,\beta) = \left(\sum_{i=1}^{n} \ln f_X(x_i;\alpha) + \sum_{i=1}^{n} \ln f_Y(y_i;\beta)\right) \tag{8}$$

$\ell_m(\alpha,\beta)$ is the log-likelihood of marginal density functions for random variables X and Y or can be called as the marginal log-likelihood model.

$$\ell_c(\alpha,\beta,\theta) = \sum_{i=1}^{n} \ln c[F_X(x_i;\alpha), F_Y(y_i;\beta);\theta] \quad (9)$$

$\ell_c(\alpha,\beta,\theta)$ is the log-likelihood of the copula density function or can be called as the copula log-likelihood model.

The steps of IFM that were used in this study are presented as below.

Step 1: The log-likelihood of the marginal distribution function (Eq. 8) is maximized to estimate the estimators of α and β.

$$\hat{\alpha},\hat{\beta} = \text{argmax } (\ell_m(\alpha,\beta)) \quad (10)$$

Step 2: α and β in the log-likelihood copula model, $\ell_c(\alpha,\beta,\theta)$ are replaced with $\hat{\alpha}$ and $\hat{\beta}$. Then, $\ell_c(\hat{\alpha},\hat{\beta},\theta)$ is maximized to estimate the dependence estimator, $\hat{\theta}$.

$$\hat{\theta} = \text{argmax } (\ell_c(\hat{\alpha},\hat{\beta},\theta)) \quad (11)$$

2.5 Adaptive Maximization by Parts

For this research, the adaptive maximization by parts (AMBP) proposed by Zhang *et al*. [4] was applied to estimate the copula dependence estimator.

Step 1: Estimate the initial parameters $(\alpha_1,\beta_1,\theta_1)$ using the IFM method.

$$\alpha_1,\beta_1 = \text{argmax } (\ell_m(\alpha,\beta)) \quad (12)$$

$$\theta_1 = \text{argmax } (\ell_c(\alpha_1,\beta_1,\theta)) \quad (13)$$

Step *k*:

$$\alpha_k,\beta_k = \text{argmax } (\ell(\alpha,\beta,\theta_{k-1})) \quad (14)$$

$$\theta_k = \text{argmax } (\ell_c(\alpha_k,\beta_k,\theta)) \quad (15)$$

For *k* = 2, 3, 4, …

As shown in Equation 12, the IFM estimators $(\hat{\alpha},\hat{\beta},\hat{\theta})_{IFM}$ is taken as the initial values of the parameters $(\alpha_1,\beta_1,\theta_1)$ in Step 1 for the AMBP steps. While, for the Step *k*, the θ in $\ell(\alpha,\beta,\theta)$ is replaced with θ_{k-1} and then the log-likelihood equation (14) is maximized with respect to the marginal parameters α,β to estimate the next (α_k,β_k). After that, same as Step 2 in the IFM method, α and β in the copula log-likelihood model, $\ell_c(\alpha,\beta,\theta)$ are replaced with estimators of α_k and β_k to estimate the next θ_k. As the number *k* tends to infinity, the estimator converges to the MLE of (α,β,θ).

The estimation performance of these three parametric methods is compared through the simulation and empirical studies.

2.6 Simulation Study

It is difficult to estimate the copula dependence parameter, θ and to compare the three parametric estimation methods theoretically. Therefore, a simulation study was conducted in order to achieve the objectives. In the simulation study, simulation data are generated from Clayton copula as the true copula with four different values of true copula parameter dependence that are corresponding to Kendall's tau, τ = 0.20, 0.50, 0.60, and 0.80. The relationship of Kendall's tau (τ) with the Clayton copula is shown in equation (16) below.

$$\tau = \frac{\theta}{\theta+2} \quad (16)$$

The sample sizes of the generated data are set to *n* = 50, 100, 1000, and 5000. 500 repetitions of data generation, estimation process and squared error calculation are done for each combination of different data sample size, *n* and copula dependence level, θ. The performance of the three estimation methods and the estimators' precision were compared based on the measured root mean square error (RMSE). The RMSE formula is given as follows:

$$RMSE(\hat{\theta}) = \sqrt{\sum_{i=1}^{500} \frac{(\hat{\theta}_i - \theta)^2}{500}} \quad (17)$$

where $\hat{\theta}_i$ is the estimator for the i^{th} replication, and θ is the true parameter used in the simulation.

The procedures for the simulation study are illustrated in Figure 2 as follows:

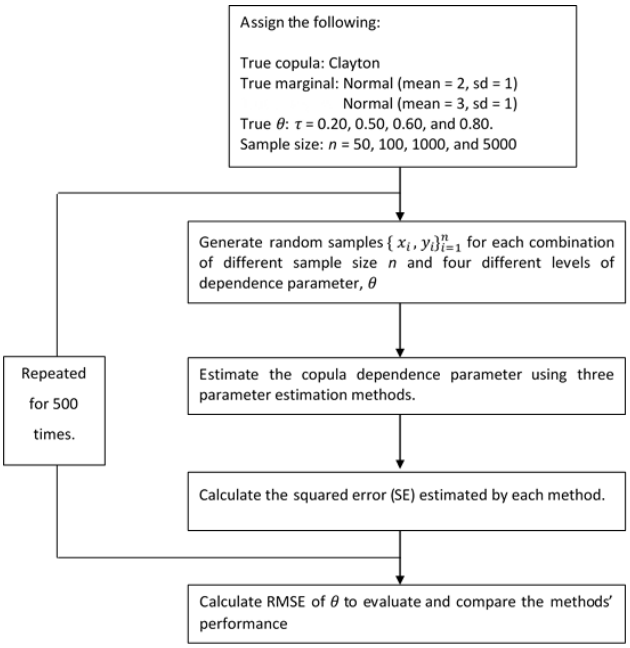


Figure 2 The procedures to compare the performance of the estimation methods in the simulation study

2.7 Empirical Study

In the empirical study, rainfall data were used in comparing the performance of the estimation methods. Three types of marginal distributions: Weibull, Gamma, and Exponential distributions are considered in fitting the hydrologic variables. This empirical study is limited only to the case of the bivariate copulas that are listed in Table 1.

Table 1 The properties of Archimedean and Elliptical copulas

Copula Family and distribution functions, $C(u,v;\theta)$	**θ range**
Clayton $(u^{-\theta} + v^{-\theta} - 1)^{-1/\theta}$	$\theta \geq -1$
Ali-Mikhail-Haq $\frac{uv}{1 - \theta(1-u)(1-v)}$	$\theta \in [-1,1]$
Frank $-\frac{1}{\theta}\ln\left[1 + \frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}\right]$	$\theta \neq 0$
Gumbel-Hougaard $\exp-\left[(-\ln u)^{\theta} + (-\ln v)^{\theta}\right]^{(1/\theta)}$	$\theta \geq 1$
Gaussian $\int_{-\infty}^{\phi^{-1}(u)}\int_{-\infty}^{\phi^{-1}(v)} \frac{1}{2\pi(1-\theta^2)^{1/2}}\exp\left\{-\frac{x^2-2xy\theta+y^2}{2(1-\theta)^2}\right\} dy\,dx$	$\theta \in [-1,1]$
Student's t $\int_{-\infty}^{t^{-1}(u)}\int_{-\infty}^{t^{-1}(v)} \frac{1}{2\pi(1-\theta^2)^{1/2}}\left\{1+\frac{x^2-2xy\theta+y^2}{(1-\theta)^2}\right\}^{-(r+1)/2} dy\,dx$	$\theta \in [-1,1]$

The empirical study was conducted by the following procedures:

Step 1 : Measure the dependency of the bivariate rainfall data in order to see the significance of the correlation and to check whether all the copula models listed in Table 1 can be used to model the dependency of the bivariate hydrologic data.

Step 2 : Fit the bivariate hydrologic data with the choice of the marginal distributions through the goodness of fit test.

Step 3 : Model the dependency of the bivariate hydrologic data by using the bivariate copulas that have been downsized from Step 1.

Step 4 : Apply the three parameter estimation methods to estimate the copula dependency parameter, θ.

Step 5 : Assess the performance of the estimation methods and identify the best-fitted copula model through the goodness of fit test.

2.8 Goodness of Fit (GOF) Test

To select a fitted marginal distribution, the statistical goodness of fit (GOF) test was applied to the empirical study in this research. GOF test is a common method to verify the fitness of the statistical model to a set of observations. The best fitted marginal and copula distribution for this research were chosen based on the smallest value Akaike Information Criterion (AIC).

The formula of AIC is written as:

$$AIC = 2p - 2\ln L \tag{18}$$

where L is the value of the likelihood function based on the estimated parameters and p is the number of estimated parameters in the statistical model.

As this study is mainly interested in the estimation of copula dependence parameter θ, the AIC values can be obtained by calculating the maximum likelihood of the copula log-likelihood model in equation (9) instead of using the full log-likelihood function in equation (7). Therefore, for copula GOF test, the formula of AIC can be expressed as:

$$AIC = 2p - 2\sum_{i=1}^{n}\ln c[\hat{F}_X(x_i), \hat{F}_Y(y_i)\,;\,\hat{\theta}] \tag{19}$$

where p is the number of parameters in the copula model, $\hat{F}_X(x_i)$ and $\hat{F}_Y(y_i)$ are the values of the estimated cumulative distribution at x_i and y_i respectively, and $\hat{\theta}$ is the estimated copula dependence parameter.

3.0 RESULTS AND DISCUSSION

The estimation performance of ML estimation, IFM and AMBP methods were compared and evaluated in the simulation study based on the RMSE value. The three parametric estimation methods were then applied to the rainfall data in Station A and Station B to estimate the dependency between them.

3.1 Simulation Study

The root mean squared errors (RMSE) for θ estimated by each method correspond to the sample size n = 50, n = 100, n = 1000 and n = 5000 are presented in Table 2. The rank of each method are based on the measured RMSE and illustrated in Figure 3. Rank 1 indicates that the method has the smallest RMSE which means the method has the best performance in parameter estimation.

Table 2 shows the RMSE for θ estimated by each method corresponding to the sample size n = 50, n = 100, n = 1000 and n = 5000. For sample size n = 50, IFM method shows higher precision with small RMSEs, giving the smallest RMSE under all correlation levels, Kendall's τ of 0.20, 0.50, 0.60, and 0.80. Whereas, the rank for AMBP and MLE is not consistent but they do give similar RMSE. For sample size n = 100. IFM method

shows higher precision with smaller RMSEs when τ = 0.20, 0.50, and 0.60 but for τ = 0.80, AMBP method has the smallest RMSE followed by MLE and IFM. For sample size *n* = 1000, it can be seen that AMBP has overtaken the ranking by showing the higher precision with small RMSEs for all correlation levels. The ranking is followed by MLE and IFM. Lastly, for sample size *n* = 5000. The results show that AMBP has the smallest RMSEs for all correlation levels. The ranking is followed by MLE and IFM.

Overall, the performance of the parametric estimation methods is different based on the sample size and the correlation level. When the sample is small, where *n* = 50, IFM method gives more precise estimator than MLE and AMBP for all correlation levels. For sample size *n* = 100, for τ = 0.2, 0.5 and 0.6, IFM performs better than MLE and AMBP. But when the correlation is very high, τ = 0.80, AMBP and MLE methods estimate more precise estimator than IFM. This is because IFM has lost the efficiency in estimation because the first step in the IFM method only considers marginal parameters but disregards the dependence level that may exist between the marginal random variables. While, for a large sample, *n* =1000 and 5000, AMBP and MLE methods estimate more precise estimator than IFM method for all correlation levels.

Therefore, based on the results of the simulation study, it can be said that for small sample size, $n <$ 1000, IFM estimator is more precise than AMBP and MLE estimators for, $\tau <$ 0.80.However, for $\tau \geq$ 0.80, MBP estimator is more precise than MLE and IFM estimators. While for large sample size, $n \geq$ 1000, MBP estimator is more precise than MLE and IFM estimators for any correlation level.

The difference between the RMSE of AMBP and MLE estimators is very small since the AMBP estimator $\hat{\theta}_{AMBP}$ converged to MLE estimator $\hat{\theta}_{MLE}$ as the iteration *k* in Step *k* in AMBP algorithm tends to infinity. However, AMBP performs better than MLE because the AMBP estimator is updated until the smallest RMSE computed, where $\hat{\theta}_{AMBP}$ converged to a constant value. Therefore, from the above results, it can be concluded that all the parametric methods could have the same performance when the sample size is large although the correlation level is small.

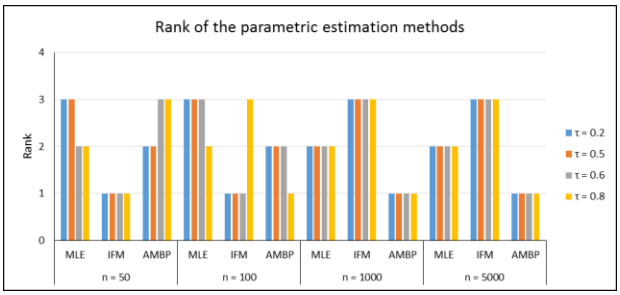


Figure 3 The ranking for each parameter estimation method based on the RMSE from Table 2

Table 2 Rank of the parametric estimation methods based on the RMSE of θ

Sample size, n	Method	τ = 0.2		τ = 0.5		τ = 0.6		τ = 0.8	
		RMSE	Rank	RMSE	Rank	RMSE	Rank	RMSE	Rank
50	MLE	0.260940	3	0.568561	3	0.761583	2	1.804203	2
	IFM	0.251549	1	0.538741	1	0.725071	1	1.669725	1
	AMBP	0.260926	2	0.568558	2	0.761706	3	1.806007	3
100	MLE	0.182119	3	0.383016	3	0.511504	3	1.194508	2
	IFM	0.178910	1	0.379302	1	0.503107	1	1.208958	3
	AMBP	0.182053	2	0.382984	2	0.511315	2	1.189954	1
1000	MLE	0.051894	2	0.117036	2	0.152343	2	0.370964	2
	IFM	0.051923	3	0.118020	3	0.153020	3	0.376295	3
	AMBP	0.051885	1	0.116982	1	0.152218	1	0.370962	1
5000	MLE	0.024233	2	0.050937	2	0.075162	2	0.158973	2
	IFM	0.024243	3	0.051235	3	0.075954	3	0.160382	3
	AMBP	0.024215	1	0.050811	1	0.074592	1	0.158784	1

3.2 Empirical Study

In this section, the three parametric copula estimation methods were applied and compared for a joint distribution identification of the rainfall data. The rainfall data used in this study is selected from two Kelantan rain gauge stations, Station Kuala Krai, 5522047 (Station A) and Station Ulu Sekor, 5520001 (Station B). Their descriptive statistics are presented in Table 3.

Table 3 Descriptive statistics of the daily rainfall for Station A and Station B

Descriptive statistics	Station A: Station Kuala Krai, 5522047	Station B: Station Ulu Sekor, 5520001
Minimum (mm)	0.0	0.0
Maximum (mm)	305.5	391.0
Mean (mm)	9.4	7.9
Standard Deviation (mm)	16.6	17.7
Coefficient of variation (CV)	177.3%	223.6%

Table 3 shows the descriptive summary measures for the daily rainfall data from Station A and Station B. The measures include the minimum, maximum, mean, standard deviation, and coefficient of variation (CV) of the rainfall data. The minimum daily rainfall recorded for both stations is 0 mm which means there was no rain on that day. The highest daily rainfall recorded is 305.5 mm at Station A and 391.0 mm at Station B. The rainfall data from Station A is averaged at 9.4 mm with standard deviation 16.6 mm. Meanwhile, at Station B, the rainfall data is averaged at 7.9 mm with standard deviation 17.7. The CV shows the dispersion of the daily rainfall and it is expressed as a percentage. The CV of daily rainfall at Station A is 177.3%, which is smaller than the CV of daily rainfall at Station B, 223.6%.

3.2.1 The Correlation Level between the Rainfalls Data

The correlation between the rainfall data from Station A and Station B is shown in the scatter plot as follows.

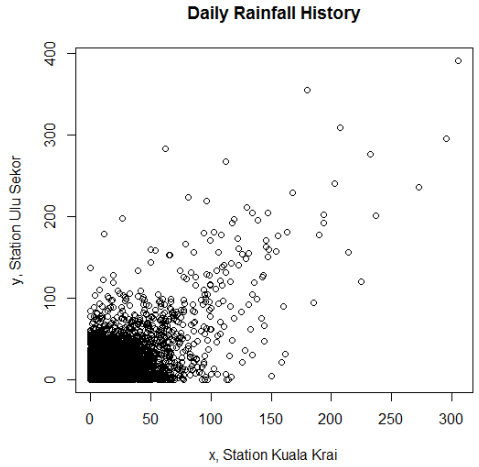


Figure 4 Scatter plot of the daily rainfall data from Station A and B in millimeter unit (mm)

It is observed that in Figure 4, the rainfall data from Station A and Station B are positively correlated. The correlation between the rainfall data Station A and Station B was measured first using Kendall's tau method. The correlation of the two series is 0.4137 with *p*-values equal to 0.000 at the significance level of α = 0.05. Since the *p*-value is less than 0.05, this means that the correlation for the rainfall data is significant.

Since the true copula and the copula dependence parameter, θ are unknown, the measured Kendall's tau can also be used to downsize the copula selection. From Kendall's tau measurement of the two stations, only five from six copulas listed in Table 1 are suitable to model the dependence between Station A and B. The five copulas are Gumbel-Hougaard, Clayton, Frank, Gaussian and Student's *t* copulas. The Ali-Mikhail-Haq copula is not considered because Kendall's tau of the two series is 0.4137 which is out of Kendall's τ range of Ali-Mikhail-Haq copula, $\tau \in$ [-0.1817, 0.3333].

3.2.2 Marginal Distributions of the Daily Rainfall Data

In applying the copula parametric estimation methods to real hydrological data, the marginal distributions need to be identified first in order to avoid the misspecification of the marginal distributions. Three types of distributions were considered in fitting the daily rainfall data: Gamma, Weibull, and Exponential. In this study, the best-fitted marginal distributions were selected based on the goodness of fit test using the Akaike Information Criterion (AIC) measurement. The parameters of the fitted marginal distribution are estimated by using maximum likelihood estimation (MLE).

In the daily rainfall data, there are some days where it did not rain and recorded as zero. Therefore, a modification has been done, where the zero values are replaced by 0.0001 in order to do the log transformation of the rainfall data. The log transformation is needed for the AIC calculation and for the steps in the MLE of marginal parameters. The goodness of fit test based on the results of AIC values is displayed in Table 4. It indicates that Gamma distribution is the best-fitted model for the daily rainfall data of both stations since the AIC values for Gamma distribution are the smallest for both rain gauge stations.

Table 4 Test of goodness-of-fit for marginal distribution based on the AIC result

Marginal Distribution	Station A: Station Kuala Krai, 5522047	Station B: Station Ulu Sekor, 5520001
	AIC	AIC
Gamma	65677.14	40538.01
Weibull	69647.91	43934.22
Exponential	105653.10	100110.80

3.2.3 Joint Daily Rainfall Data by Copula Method

The following copula estimation is then carried out for the daily rainfall data from the two stations. Gamma distribution is used as the marginal distributions for the parametric estimation methods: MLE, IFM, and AMBP since these methods need the marginal information. For Station A, the estimated shape parameter is α = 0.2522 and the scale parameter is β = 37.1838. While, for Station B, the estimated shape parameter is α = 0.2043 and the scale parameter is β = 38.7087.

The copula distribution that can describe the relationship between rainfalls for both station is still unknown. Thus, we need to have a list of suitable copula candidates. In this study, the copula candidates selected after they have been downsized were applied to model the dependence of daily rainfall at the two rain gauge stations. There are three copula models under Archimedean which are Gumbel-Hougaard, Clayton, and Frank copula and two elliptical copula families, Gaussian and Student's *t* copulas. The estimated dependence parameter of the five candidate copulas using the three estimation methods are given in Table 5. It can be seen that the AMBP estimator is very close to MLE estimator. This is because in the AMBP algorithm, the estimator for $(\alpha_k, \beta_k, \theta_k)$ is constant for each *k* iteration, and as the number *k* tends to infinity, the estimator converge to the MLE of (α, β, θ). This result is consistent with the findings from Song *et al*. [7] and Zhang *et al*. [4].

Table 5 The estimators of the dependence parameter

Copula	MLE	IFM	AMBP
Gumbel	1.7029	1.688	1.7029
Clayton	0.834	0.8241	0.834
Frank	4.0959	4.1147	4.0959
Gaussian	0.5522	0.5577	0.5522
Student's t	0.6006	0.5815	0.6005
df	1.7365	2.3789	1.7375

#df = the estimator for degree of freedom for Student's copula.

To select a fitted copula model and to measure the performance of estimation methods, the statistical goodness of fit (GOF) test has been applied for the empirical study in this research. The GOF test describes the fitness of the model to a set of observations. The best-fitted distribution is determined based on the minimum error produced, which is measured by Akaike Information Criterion (AIC) for this study.

A small AIC value represents a better model fit. The AIC of each copula estimated by different estimation methods are listed in Table 6.

Table 6 Test of goodness-of-fit for copula function based on the AIC result

Copula	MLE	IFM	AMBP
Gumbel	-7766.77	-7730.40	-7766.77
Clayton	-5554.23	-5463.15	-5554.23
Frank	-6613.32	-6600.97	-6613.32
Gaussian	-6670.03	-6611.80	-6670.04
Student's *t*	**-8932.50**	**-8359.70**	**-8932.50**

Table 6 shows that the AIC of the Student's *t* copula estimated by MLE, IFM, and AMBP are smaller than the AIC of the other copulas. It shows that all estimation methods identify Student's *t* copula as the best one among the five candidate copulas that can describe the dependency of the rainfall data from Station A and Station B. Since the best-fitted copula has been determined, the performance of the three estimation methods can be compared based on the estimated copula estimator of Student's *t* copula and the estimated AIC.

From the results in Table 6, it can be seen that the AMBP and MLE methods have estimated the *Student's t* copula estimators, $\hat{\theta}$ that are almost similar in values, which are 0.6005 and 0.6006 respectively. However, the IFM estimator seems to have a larger difference compared to the other parametric estimators. This is because the first step in IFM method only estimates the marginal parameters without considering the correlation that exists between the rainfall variables. This empirical result is consistent with the simulation. Since the sample size for the rainfall data is very large, which is about 16314, this condition also contributes to a precise copula estimator. The AIC estimated by AMBP method is the smallest followed by the estimated AIC by MLE and IFM methods.

4.0 CONCLUSION

The simulation and empirical results from this study have given the statistical evidence in choosing which parameter estimation methods that are more accurate and efficient to estimate the copula dependence parameter. Between the parametric approaches, IFM method can estimate efficiently enough when the sample size is small, e.g., $n < 1000$ and the correlation level is less than 0.80. When the sample size is large, e.g., $n \geq 1000$ and the variables are significantly correlated, AMBP method should be used in order to estimate a precise and efficient estimator.

However, this study also has some limitations. First, this study only used normal distribution as the true marginal distribution for the simulated variables *X* and *Y*. Hence, a future study should apply different marginal distributions for random variables *X* and Y. The combination of different marginal distribution is

able to describe the advantage of using Copula method to model the joint distribution. Second, this study only focuses on the comparison of parametric approaches. Thus, this research can be extended by comparing the parametric approaches with the semiparametric and nonparametric approaches for copula parameter estimation, such as Pseudo maximum likelihood (PML), Bayesian approach or Kernel density estimation for copula. Furthermore, this research can also be improved by using other difference performance measures and goodness-of-fit tests.

Acknowledgement

This work is fully supported by Ministry of High Education (MOHE) and Universiti Teknologi Malaysia (UTM). The authors would like to thank the MOHE for the STEM Grant with vote number A. J091002.5600.07397 and the University for the Grant with vote number Q.J130000.2426.04G34 for providing the support and sponsorship.

References

[1] Sklar, A. 1959. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de L'Université de Paris*. 8: 229-231.
[2] Zhang, L., & Singh, V. P. 2007. Bivariate Rainfall Frequency Distributions Using Archimedean Copulas. *Journal of Hydrology*. 332(1-2): 93-109. DOI: https://doi.org/10.1016/j.jhydrol.2006.06.033.
[3] Yee, K. C., Jamaludin, S., Yusof, F., and Mean, F. H. 2014. Bivariate Copula in Fitting Rainfall Data. *AIP Conference Proceedings*. 1605(1): 986-990. DOI: http://dx.doi.org/10.1063/1.4887724.
[4] Kim, G., Silvapulle, M., and Silvapulle, P. 2007. Comparison of Semiparametric and Parametric Methods for Estimating Copulas. *Computational Statistics & Data Analysis*. 51: 2836-2850. DOI: https://doi.org/10.1016/j.csda.2006.10.009.
[5] Kim, T. W., Valdés, J. B., and Yoo, C. 2006. Nonparametric Approach for Bivariate Drought Characterization Using Palmer Drought Index. *Journal of Hydrologic Engineering*. 11(2): 134-143. DOI: https://doi.org/10.1061/(ASCE)1084-0699(2006)11:2(134).
[6] Kojadinovic, I., and Yan, J. 2010. Comparison of Three Semiparametric Methods for Estimating Dependence Parameters in Copula Models. *Insurance: Mathematics and Economics*. 47: 52-63. DOI: https://doi.org/10.1016/j.insmatheco.2010.03.008.
[7] Lawless, J. F., and Yilmaz, Y. E. 2011. Comparison of Semiparametric Maximum Likelihood Estimation and Two-Stage Semiparametric Estimation in Copula Models. *Computational Statistics and Data Analysis*. 55: 2446-2455. DOI: https://doi.org/10.1016/j.csda.2011.02.008.
[8] Nagler, T., Schellhase, C., and Czado, C. 2017. Nonparametric Estimation of Simplified Vine Copula Models: Comparison of Methods. *Dependence Model*. 5: 99-120. DOI: https://doi.org/10.1515/demo-2017-0007.
[9] Taheri, B. M., Jabbari, H., and Amini M. 2018. Parameter Estimation of Bivariate Distributions in Presence of Outliers: An Application to FGM Copula. *Journal of Computational and Applied Mathematics*. 343: 155-173. DOI: https://doi.org/10.1016/j.cam.2018.04.043.
[10] Dupuis, D. J. 2007. Using Copulas in Hydrology: Benefits, Cautions, and Issues. *Journal of Hydrologic Engineering*. 12(4): 381-393. DOI: https://doi.org/10.1061/(ASCE)1084-0699(2007)12:4(381).
[11] Zhang, R., Czadoa, C., and Min, A. 2011. Efficient maximum Likelihood Estimation of Copula based Meta t-distributions. *Computational Statistics & Data Analysis*. 55(3): 1196-1214. DOI: https://doi.org/10.1016/j.csda.2010.09.027.
[12] Joe, H. and Xu, J. J. 1996. The Estimation Method of Inference Functions for Margins for Multivariate Models. *Technical Report no. 166*. Department of Statistics, University of British Columbia. DOI: http://dx.doi.org/10.14288/1.0225985.
[13] Joe, H. 2005. Asymptotic Efficiency of the Two-Stage Estimation Method for Copula-Based Models. *Journal of Multivariate Analysis*. 94: 401-419. DOI: https://doi.org/10.1016/j.jmva.2004.06.003.
[14] Song, P. X. K., Fan, Y., and Kalbfleisch, J. D. 2005. Maximization by Parts in Likelihood Inference. *Journal of the American Statistical Association*. 100(472): 1145-1158. DOI: http://dx.doi.org/10.1198/016214505000000204.
[15] Silvennoinen, A. and Teräsvirta, T. 2017. Consistency and Asymptotic Normality of Maximum Likelihood Estimators of A Multiplicative Time-varying Smooth Transition Correlation GARCH Model. *Research Paper 2017-28*. CREATES, Aarhus University, Aarhus.
[16] Frazier, D. T. and Renault, E. 2017. Efficient Two-step Estimation Via Targeting. *Journal of Econometrics*. 201(2): 212-227. DOI: http://dx.doi.org/10.1016/j.jeconom.2017.08.004.
[17] Ariff, N.M., Jemain, A.A., Ibrahim, K., and Wan Zin, W.Z. 2012. IDF Relationships Using Bivariate Copula for Storm Events In Peninsular Malaysia. *Journal of Hydrology*. 470-471, 158-171. DOI: https://doi.org/10.1016/j.jhydrol.2012.08.045.
[18] Requena, A. I., Mediero, L., and Garrote, L. 2013. A Bivariate Return Period Based on Copulas For Hydrologic Dam Design: Accounting for Reservoir Routing in Risk Estimation. *Hydrology and Earth System Sciences*. 17: 3023-3038. DOI: https://doi.org/10.5194/hess-17-3023-2013.
[19] Yusof, F., Mean, F. H., Jamaludin, S., and Yusof, Z. 2013. Characterization of Drought Properties with Bivariate Copula Analysis. *Water Resource Management*. 27: 4183-4207. DOI: https://doi.org/10.1007/s11269-013-0402-4.
[20] Vandenberghe, S., Verhoest, N. E. C., and De Baets, B. 2010. Fitting Bivariate Copulas to the Dependence Structure between Storm Characteristics: A Detailed Analysis based on 105 Year 10 Min Rainfall. *Water Resources Research*. 46: W01512. DOI: http://dx.doi.org/10.1029/2009WR007857.
[21] Chen, L., Singh, V. P., Guo, S., Zhou, J., and Zhang, J. 2015. Copula-based Method for Multisite Monthly and Daily Streamflow Simulation. *Journal of Hydrology*. 528: 369-384. DOI: https://doi.org/10.1016/j.jhydrol.2015.05.018.
[22] Clayton, D. G. 1978. A Model for Association in Bivariate Life Tables and Its Application in Epidemiological Studies of Familial Tendency in Chronic Disease Incidence. *Biometrika*. 65(1): 141-151. DOI: https://doi.org/10.1093/biomet/65.1.141.