Using the skew-t copula to model bivariate rainfall distribution

R. Zakaria[1] A. V. Metcalfe[2] P. Howlett[3]
J. Piantadosi[4] J. Boland[5]

(Received 16 March 2010; revised 27 April 2010)

Abstract

We simulate monthly rainfall at two sites in the Murray–Darling Basin. In order to construct a suitable joint distribution, we model the individual totals using appropriate gamma distributions and use a multivariate skew-t distribution to construct an appropriate copula. The skew-t distribution is considered robust as it includes both skewness and tail dependence structure and allows us to model correlations. We investigate the characteristics of a bivariate skew-t distribution and show how adjusting the parameters generates simulated data which matches the observed data.

Contents

http://anziamj.austms.org.au/ojs/index.php/ANZIAMJ/article/view/2030 gives this article, © Austral. Mathematical Soc. 2010. Published May 13, 2010. ISSN 1446-8735. (Print two pages per sheet of paper.)

1 Introduction and motivation

This study is aimed at modelling monthly rainfall at two different sites within the same local region of the Murray–Darling Basin. The methodology developed here can be extended to multiple sites within the same area and the further extension will be to construct a rainfall-runoff model for a catchment. The Murray–Darling Basin is very important to Australia as it represents 39% of Australian national income from agricultural production and in 2004–2005 it consumed about 83% of the water used for agriculture [7]. Hence, extensive study needs to be conducted related to water issues such as rainfall modelling. In rainfall modelling, we try to develop a model to generate synthetic rainfall data whose statistical characteristics match those of the historical data. If done successfully, we can extend this mechanism further to develop other water related models such as a rainfall-runoff model.

Rainfall data has a skewed distribution. As well we are interested in the interdependence of rainfall events at rainfall stations. The copula method, which generates a general structure and can accommodate extreme events, has been used to model the interdependency. Copulas are functions that join multivariate distributions to their one dimensional marginal distribution functions [6]. Some examples of copula types are the Gaussian copula and the Student's t copula (t-copula). The Gaussian copula is based on a multivariate normal distribution and does not model tail dependence, whereas the t-copula is based on the multivariate Student's t distribution and does accommodate tail dependence. The tail dependence is used to define the degree of dependence in the lower or upper tails of a bivariate distribution. This concept is widely applied in finance to model dependence of loss events with various assets [4]. In this study, the asymmetric t-copula, also known as skew-t, is employed to analyse the asymmetric tail dependence using monthly rainfall data.

2 Study area and data

The method is illustrated using monthly rainfall data from two different sites within the same local region of the Murray–Darling Basin. The selected sites are Hume and Beechworth, 46 km apart, in New South Wales and Victoria respectively (Table 1). The data used was provided by the Australian Bureau of Meteorology for a continuous period from 1928 until 1985 (58 years). Table 2 shows that even though the sites are in close proximity, Beechworth receives significantly greater rainfall on average, but also with greater variability.

TABLE 1: Description of two selected sites in Murray–Darling Basin.

Station name	State	Abbreviation	Latitude	Longitude
Hume Reservoir	NSW	Hume	−36.1039	147.0329
Beechworth Composite	VIC	Beech	−36.3702	146.7132

TABLE 2: Summary of rainfall mean (mm) and standard deviation (SD) (mm).

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Hume												
Mean	43	43	46	55	61	65	74	78	62	74	51	49
SD	45	44	39	41	44	41	39	39	32	45	37	47
Beech												
Mean	52	52	62	80	93	101	112	117	93	101	68	61
SD	46	58	47	59	66	63	56	53	46	52	42	47

3 Rainfall model

We discuss the theory of the development of our rainfall model. We choose a skew-t copula based on the reasons mentioned in Section 1. The theory of the univariate skew-t followed by the multivariate skew-t distribution are presented in this section.

3.1 Univariate skew-t distribution

A random variable X has a skew-t distribution, $X \sim ST(\mu, \omega, \gamma, \nu)$ if it has a probability density function [3]

$$f_{ST}(x; \mu, \omega, \gamma, \nu) = \frac{2}{\omega}\, t(z; \nu)\, T\left\{\gamma z \sqrt{\frac{\nu+1}{\nu+z^2}}; \nu+1\right\}, \quad x \in \mathbb{R},$$

where $z = (x - \mu)/\omega$, $t(\cdot)$ and $T(\cdot)$ stand for the univariate standard Student's t probability density function and cumulative distribution function with $\nu + 1$, respectively, with location parameter $\mu \in \mathbb{R}$, scale parameter $\omega \in (0, \infty)$, shape parameter $\gamma \in \mathbb{R}$ and degrees of freedom ν. Skewness and tail proportions are controlled by parameters γ and ν, respectively. When $\gamma = 0$, the standard t distribution is obtained and when $\nu \to \infty$, the skew normal distribution is formed. However when both $\gamma = 0$ and $\nu \to \infty$, the normal distribution is recovered.

3.2 Multivariate skew-t distribution

A higher dimension of skew-t distribution to handle multivariate data is written as $MST(\mu, \Omega, \gamma, \nu)$ which has a probability density function of $x \in \mathbb{R}^d$ [2],

$$f_{MST}(x; \mu, \Omega, \gamma, \nu) = 2t_d(x - \mu; \Omega, \nu)\ T\left\{\gamma^T \Omega^{-1}(x - \mu)\sqrt{\frac{\nu + d}{Q(x) + \nu}}; \nu + d\right\},$$

where $Q(x) = (x-\mu)^T\Omega^{-1}(x-\mu)$. The d-dimensional Student's t distribution with zero location ($\mu = 0$), Ω scale matrix and ν degrees of freedom is defined as

$$t_d(x; \Omega, \nu) = \frac{\Gamma[(\nu + d)/2]}{\sqrt{|\Omega|}(\nu\pi)^{d/2}\Gamma(\nu/2)}\left(1 + \frac{Q(x)}{\nu}\right)^{-(\nu+d)/2}, \quad x \in \mathbb{R}^d,$$

where $T(\cdot)$ is the cumulative distribution function of Student's t distribution with $\nu + d$ degrees of freedom.

4 Simulation experiment

As an illustration for the simulation experiment, we use a bivariate skew-t distribution. The procedure is divided into three parts:

1. fitting data to gamma marginals,
2. generating simulated data using the skew-t distribution and
3. comparing the analysis from Part 1 and Part 2.

4.1 Fitting data to gamma marginals

We select two sets of monthly rainfall data (1928–1985, 58 years) from two adjacent rainfall stations in the Murray–Darling Basin, Hume and Beechworth. The stations are 46 km apart with lag zero cross-correlation of 0.88. The rainfall data is categorised into four different cases: (wet,wet), (wet,dry), (dry,wet) and (dry,dry). For illustration, we only consider the first case, that is (wet,wet) when both stations have rain. Therefore, we have pairs of positive numbers. The individual sets of rainfall data are fitted separately using gamma marginals. The gamma distribution is chosen as it is suitable to model continuous variables that are always positive and have a skewed distribution like rainfall totals. Furthermore, it is also flexible as it involves two parameters, scale and shape. The probability density function for the gamma distribution is

$$f(x, \alpha, \beta) = \frac{x^{\alpha-1} e^{-x/\beta}}{\Gamma(\alpha)\beta^{\alpha}}, \quad x > 0 \text{ and } \alpha, \beta > 0.$$

The shape parameter α and the scale parameter β for each month are estimated using the maximum likelihood method (Table 3). These sets of monthly rainfall data between Hume and Beechworth are then transformed into bivariate uniform data using the fitted cumulative gamma density function.

The correlation and the tail proportion of the uniform data for Hume and Beechworth are calculated (Table 4). In the following section we generate simulated data using the bivariate skew-t.

TABLE 3: The monthly values of shape parameter α and scale parameter β.

	Hume		Beechworth	
	α	β	α	β
Jan	0.90	48.39	1.22	42.89
Feb	1.01	44.26	1.07	49.11
Mar	1.24	37.34	1.40	44.38
Apr	1.85	29.93	1.87	42.97
May	1.85	33.50	2.06	46.07
Jun	2.45	26.54	2.74	36.75
Jul	3.67	20.26	4.44	25.28
Aug	3.02	25.73	3.46	33.80
Sep	3.06	20.14	3.75	24.76
Oct	2.33	31.57	3.24	31.11
Nov	1.86	27.38	2.12	32.24
Dec	0.96	50.32	1.52	40.21

TABLE 4: Tail proportions and correlation coefficient.

	Tail proportions		Correlation (r)
	lower 10%	upper 10%	
Observed	0.081	0.056	0.893

4.2 Generation of simulated data using skew-t distribution

For simulation we used codes from the 'sn' package in R developed by [1]. The 'sn' stands for skew normal and originally the codes were developed for the skew normal distribution. Then, the author included the codes for the skew-t distribution in the same package. The codes used are probability density function, distribution function and random number generation for the skew-t distribution (univariate: dst, pst, rst; multivariate: pmst, dmst, rmst). We generated 10^5 simulated data values for the bivariate skew-t using random number generation of multivariate skew-t, rmst$(n, \mu_i, \Omega, \gamma_i, \nu_i)$ for $i = 1, 2$. The mean and covariance were set to $\mu_1 = \mu_2 = 0$, $\omega_{11} = \omega_{22} = 1$ and $\omega_{12} = \omega_{21} = 0$. These values can be chosen arbitrarily as the effects are removed when we transform the data to uniform. Therefore, there are three free parameters namely; skewness γ, correlation ρ and degrees of freedom ν. By varying the free parameters, we investigated the effect on correlation r and tail proportions of the lower and upper 10% of the simulated data. The main objective of this exercise is to match the correlation coefficient of the observed data with that of the simulated data. The set of random numbers generated by the bivariate skew-t can be between $-\infty$ and $+\infty$. Each set of bivariate skew-t data generated was transformed to uniform data using marginals of univariate skew-t, pst$(n, \mu = 0, \omega = 1, \gamma, \nu)$. Then, the correlation r and the tail proportions (lower 10% and upper 10%) for the simulated uniform data were calculated.

4.3 Analysis of the simulation process

The analysis of the simulation process is divided into two parts as follows.

4.3.1 The effect of parameter variation on correlation and tail proportions of simulated data

In each simulation, we generated 10^5 simulated data values. Firstly, the values of parameters were varied systematically and we tried various combinations such as increasing the skewness value but with other parameters fixed. Then we did the same by increasing the degrees of freedom and fix other parameters. In each iteration, the value of simulated correlation and the tail proportions were noted. Based on the simulated results we finally decided that we should focus on finding the best values for the simulated parameters by varying the values of skewness γ from 0 to 2, degrees of freedom ν from 3.1 to infinity and model correlation ρ from 0.8 to 0.99. Varying the degrees of freedom has only a slight effect on the simulation results (Table 6). Therefore, we concentrated on variation of the skewness and correlation parameters. Increasing the skewness, decreases the correlation of the uniformised simulated data (Table 5). In order to match the correlation of the uniformised simulated data with the uniformised observed data, the correlation ρ chosen for the simulated data needs to be as high as possible where $0.9 < \rho < 1$. Note that ρ is the model correlation which we used to generate synthetic data with correlation r that matches the observed correlation. Table 5 shows that as the value of skewness increases, the estimated values for the correlation decreases. However, there is little effect on the value of the estimated correlation when the degrees of freedom varies (Table 6). The correlation for simulated data increases as the correlation ρ increases (Table 7).

4.3.2 Refining the simulation result

We would like to match the statistics of the uniformised simulated data with the uniformised observed data as closely as possible. Based on the analysis of simulation results in the previous section, the simulation results were refined further by repeating the same process by carefully choosing the right combi-

Table 5: Variation of skewness γ.

Model parameters			Simulation results		
ρ	γ	ν	r	lower 10%	upper 10%
0.89	0.0	5	0.8721	0.0696	0.0696
	0.5		0.8273	0.0594	0.0679
	1.0		0.7699	0.0471	0.0670
	1.5		0.7262	0.0368	0.0665
	2.0		0.7091	0.0292	0.0668

Table 6: Variation of degrees of freedom ν.

Model parameters			Simulation results		
ρ	γ	ν	r	lower 10%	upper 10%
0.89	1.5	3.1	0.7389	0.0381	0.0669
		3.5	0.7252	0.0358	0.0700
		4	0.7304	0.0370	0.0678
		5	0.7299	0.0371	0.0669
		10	0.7281	0.0362	0.0643
		Inf	0.7270	0.0359	0.0627

Table 7: Variation of model correlation ρ.

Model parameters			Simulation results		
ρ	γ	ν	r	lower 10%	upper 10%
0.80	1.5	5	0.556	0.023	0.056
0.85			0.651	0.029	0.066
0.90			0.752	0.039	0.069
0.95			0.867	0.055	0.077
0.98			0.945	0.069	0.086
0.99			0.971	0.078	0.090

nation of the parameters. Therefore, the simulation process was reiterated to refine the simulation results with a sample size of 10^4. Each simulation was repeated five times and the mean of the results calculated. The set of parameters that best match the observed statistics are $\rho = 0.965$, $\gamma_1 = \gamma_2 = -1.95$ and $\nu = 5$. These parameters give the simulated correlation $r = 0.892$ and tail proportions of 0.082 and 0.055, for lower 10% and upper 10% respectively (Table 8). In comparison, the statistics of the observed data are correlation 0.893 and with tail proportions of 0.081 and 0.056, for lower 10% and upper 10%, respectively.

5 Goodness of fit test

We use a graphical method to evaluate the goodness of fit between the observed and the generated data. Figure 1 presents a scatter plot of empirical copula versus theoretical copula as has been used by Genest and Favre [8] and Wong et al. [5]. The empirical copula C_n is constructed using the observed data which is transformed to uniform data on $[0, 1]$ using the gamma marginals as described in Section 4.1. Then the uniform data ranked and used to calculate C_n. For the theoretical copula $\widetilde{C}$, we generate 10^4 bivariate data values from the bivariate skew-t model which is then transformed to uniform using the univariate skew-t for each marginal. The generated uniform data is also ranked and used to calculate $\widetilde{C}$. The empirical copula is

$$C_n(u, v) = \frac{1}{n} \sum_{i=1}^{n} I\left(\frac{R_i}{n+1} \leqslant u, \frac{S_i}{n+1} \leqslant v\right),$$

where n is the sample size, R_i and S_i are the rank for each marginal. In our example we only deal with bivariate data; however, the formula can be easily extended to multivariate data. $I(A)$ defines the indicator variable for a logical expression A: that is, $I(A) = 1$ if A is true; and $I(A) = 0$ if A is false. The axes are rescaled on $[0, 1]^2$ and a set of points within that configuration

is calculated. For example, the point $(0.025, 0.023)$ on the scatter plot of theoretical versus empirical copula (Figure 1) is calculated from

$$\frac{1}{691}\sum_{i=1}^{691} I\left(\frac{R_i}{691+1} \leqslant 0.2, \frac{S_i}{691+1} \leqslant 0.3\right)$$

and

$$\frac{1}{10000}\sum_{i=1}^{10000} I\left(\frac{R_i}{10000+1} \leqslant 0.2, \frac{S_i}{10000+1} \leqslant 0.3\right).$$

The fitness of the empirical and theoretical copula is based on the closeness of the points to the line $y = x$. However, the Kolmogorov–Smirnov test is also conducted to compare the distributions of the empirical and theoretical copula. The p-value for the test of equality of distributions is 0.967 which is greater than 0.05 significance level, therefore we conclude that the two distributions are not significantly different from one another.

6 Conclusion

The bivariate skew-t distribution is presented and the application of the theory is also examined using monthly rainfall data from two sites in the Murray–Darling Basin. Each marginal distribution of the observed data set is transformed to a uniform distribution on $[0, 1]$ using the gamma distribution. The simulated data are generated using the bivariate skew-t and each marginal distribution of the generated data set is transformed to a uniform distribution using the univariate skew-t. A simulation approach estimates the parameters of the skew-t distribution. A refining process matches the statistics between the observed and simulated data which is based on their correlation coefficient and tail proportions. A plot of empirical copula versus theoretical copula illustrates the goodness of fit test coupled with the Kolmogorov–Smirnov test. The results obtained demonstrate that the skew-t copula is suitable for modelling monthly rainfall totals for correlated stations.

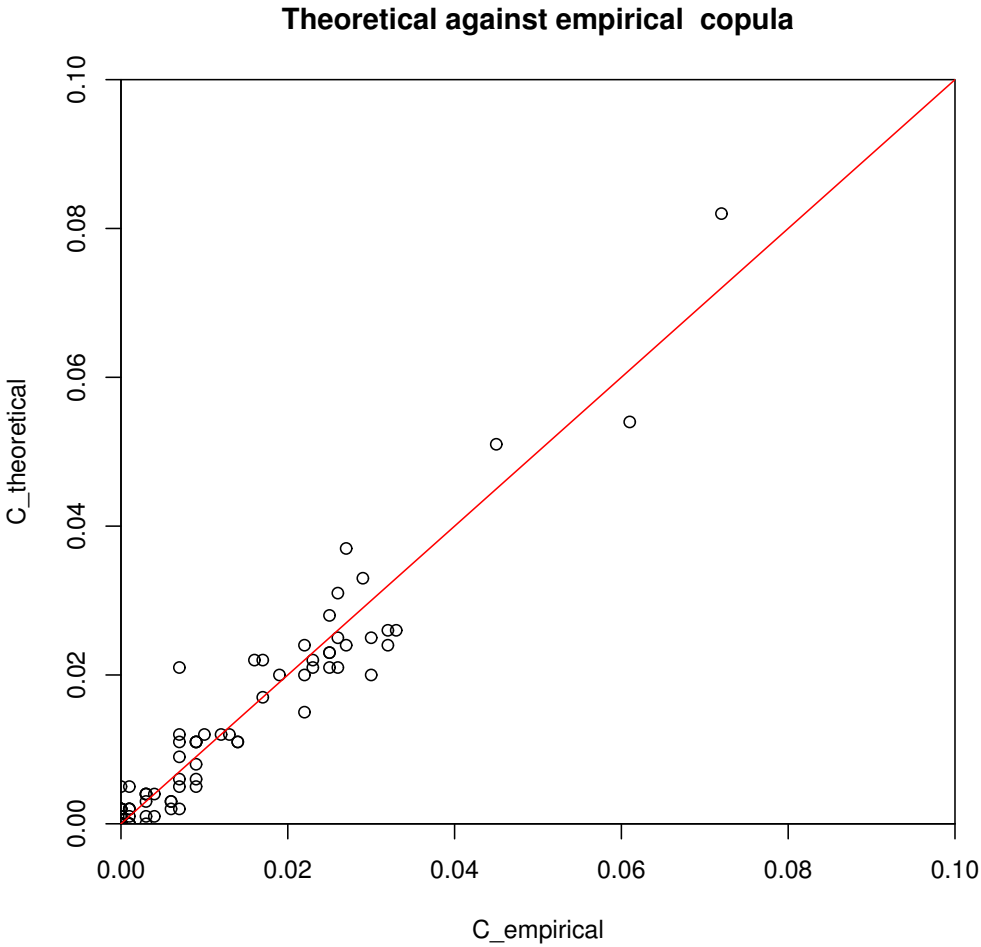

FIGURE 1: Comparison of the theoretical and empirical copulas.

TABLE 8: Refining simulation process.

Model parameters			Simulation results		
ρ	γ	ν	r	lower 10%	upper 10%
0.90	2.0	5	0.733	0.031	0.065
	−0.1	5	0.881	0.069	0.070
	−2.0	5	0.732	0.069	0.032
0.91	−0.1	5	0.894	0.074	0.071
	−1.0	5	0.802	0.068	0.050
	−1.5	5	0.767	0.070	0.041
0.92	−0.1	5	0.906	0.075	0.074
	−0.3	5	0.891	0.073	0.067
	−1.5	5	0.790	0.071	0.045
0.93	−0.1	5	0.914	0.079	0.073
	−0.4	5	0.895	0.074	0.069
	−1.5	5	0.809	0.073	0.043
0.94	−0.1	5	0.928	0.082	0.078
	−0.6	5	0.891	0.073	0.065
	−1.5	5	0.842	0.075	0.051
0.95	−0.3	5	0.932	0.080	0.079
	−0.8	5	0.898	0.075	0.065
	−2.0	5	0.857	0.082	0.049
	−2.1	5	0.849	0.077	0.046
0.96	−0.3	5	0.946	0.082	0.082
	−1.3	5	0.895	0.076	0.064
	−2.0	5	0.880	0.079	0.053
0.965	−1.0	5	0.917	0.084	0.070
	−1.5	5	0.904	0.080	0.061
	−1.8	5	0.895	0.081	0.054
	−1.9	5	0.892	0.082	0.053
	−1.95	5	0.892	0.082	0.055
observed values			0.893	0.081	0.056

Acknowledgements This research was supported by the Australian Research Council Discovery Project (DP0877707). The open source software R performed the data manipulation. We are grateful to Chris Brien for helping us with the programming.

References

[1] A. Azzalini. *R package sn: The skew-normal and skew-*t *distributions (version 0.4-12)*. Università di Padova, Italia, 2009. http://azzalini.stat.unipd.it/SN. C238

[2] A. Azzalini and A. Capitanio. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society: Series B*, 65(2):367–389, 2003. http://www.blackwell-synergy.com/doi/abs/10.1111/1467-9868.00391. C235

[3] A. Azzalini and M. G. Genton. Robust likelihood methods based on the skew-t and related distributions. *International Statistical Review*, 76(1):106–129, 2008. http://dx.doi.org/10.1111/j.1751-5823.2007.00016.x. C234

[4] P. Embrechts, F. Lindskog, and A. Mc Neil. *Modelling Dependence with Copulas and Applications to Risk Management In: Handbook of Heavy Tailed Distributions in Finance*, chapter 8, pages 329–384. Elsevier, 2003. C233

[5] C. Genest and A.-C. Favre. Everything you always wanted to know about copula modelling but were afraid to ask. *Journal of Hydrologic Engineering*, 12(4):347–368, 2007. http://dx.doi.org/10.1061/(ASCE)1084-0699(2007)12:4(347). C241

[6] R. B. Nelsen. *An Introduction to Copulas*. Springer, New York, 2nd edition, 2006. C233

[7] Murray-Darling Basin Authority. http://www.mdba.gov.au, 2008. [accessed 10-Jan-2010]. C232

[8] G. Wong, M. F. Lambert, M. Leonard, and A. Metcalfe. Drought analysis using trivariate copulas conditional on climatic states. *Journal of Hydrologic Engineering*, 15(2):129–141, 2010. http://dx.doi.org/10.1061/(ASCE)HE.1943-5584.0000169. C241

Author addresses

1. **R. Zakaria**, School of Mathematics and Statistics, University of South Australia, Adelaide, South Australia, Australia.
2. **A. V. Metcalfe**, School of Mathematics and Statistics, University of Adelaide, Adelaide, South Australia, Australia.
3. **P. Howlett**, School of Mathematics and Statistics, University of South Australia, Adelaide, South Australia, Australia.
4. **J. Piantadosi**, School of Mathematics and Statistics, University of South Australia, Adelaide, South Australia, Australia.
5. **J. Boland**, School of Mathematics and Statistics, University of South Australia, Adelaide, South Australia, Australia.