

Chapter 4

Simulation

4.1 Overview

In this chapter, a Monte Carlo simulation experiment is undertaken to test the modelling properties of the mixture copula as compared to that of the unmixed, parent copula. To proceed, data is generated from a parent copula and a mixture copula respectively. Then, for each set of data, the parent and mixture families of copulas are fitted, with the aim being to find whether, and when, either copula has modelling advantage over the other. The two copula families which will be examined in particular are the AMH family (2.8) and its *Beta*-mixture counterpart with fixed β parameter (3.17).

For simplicity, bivariate distributions with standard uniform (*Uniform*(0, 1)) marginal distributions will be considered. In that case, the copula of the joint distribution is equivalent to the joint cdf.

4.2 Simulation algorithms

A number of methods are available to generate pseudorandom data from a given copula-based distribution. One such method is the Conditional Distribution Method, which separates the copula into several univariate components, each of which can be easily sampled (Johnson [1987]). Another method, used for Archimedean copulas, uses the generator function to calculate random variates from uniform distributions (Genest and McKay [1986]).

In general, to generate a pseudorandom drawing from the m -variate $X = (X_1, X_2, X_3 \dots X_m)'$ by the Conditional Distribution Method, the procedure is as follows:

1. Generate a draw $X_1 = x_1$ from the marginal distribution of X_1 ;

2. Generate a draw $X_2 = x_2$ from the conditional distribution of X_2 , given that $X_1 = x_1$;
3. Generate a draw $X_3 = x_3$ from the conditional distribution of X_3 , given that $X_1 = x_1$ and $X_2 = x_2$,

and so on, for all X_m .

Consider, then, a bivariate distribution formed by a 2-copula with uniform margins. For this bivariate distribution, the procedure is then to generate the observation of one margin, say U , from its marginal distribution, and then to generate an observation for V from its conditional distribution given U . Now, the marginal distribution of U is *Uniform*(0,1) as defined above. The conditional distribution for V , given $U = u$, is (Nelsen [2006]):

$$c_u(v) = \Pr[V \leq v | U \leq u] = \frac{\partial}{\partial u} C(u, v).$$

Thus, to sample (u, v) from a uniform-margined, copula-based distribution $C(u, v)$, the procedure is as follows:

1. Generate two independent *Uniform*(0,1) observations u and t ;
2. Find the conditional distribution $c_u(v) = \frac{\partial}{\partial u} C(u, v)$; find its quasi-inverse $c_u^{(-1)}(t)$; set $v = c_u^{(-1)}(t)$;
3. The pair (u, v) are the sample required.

Now, consider the AMH copula, which, for $u, v \in \mathbb{I}$ and $-1 \leq \theta < 1$, has the form, from (2.8):

$$C_\theta(u, v) = \frac{uv}{1 - \theta(1-u)(1-v)}.$$

Firstly, the conditional distribution

$$\begin{aligned} c_u(v) &= \frac{\partial}{\partial u} C_\theta(u, v) \\ &= \frac{v[(1 - \theta(1 - v))]}{[1 - \theta(1 - u)(1 - v)]^2}. \end{aligned}$$

To find its inverse $c_u^{(-1)}(t)$, set $c_u(v) = t$, which leads to the following equation:

$$v[1 - \theta(1 - v)] - [1 - \theta(1 - u)(1 - v)]^2 t = 0.$$

Set $a = 1 - u$ and rearranging, we obtain the following quadratic in v :

$$(\theta - a^2\theta^2)v^2 + [-\theta(2at + 1) + 2a^2\theta^2t + 1]v + (-t + 2a\theta t - a^2\theta^2t) = 0.$$

Now, using the quadratic formula to solve for v , we obtain:

$$v = \frac{-[\theta(2at + 1) + 2a^2\theta^2t + 1] \pm \sqrt{\theta^2(4a^2t - 4at + 1) - \theta(4at - 4t + 2) + 1}}{2(\theta - a^2\theta^2)}.$$

Now set:

$$\begin{aligned} b &= -\theta(2at + 1) + 2a^2\theta^2t + 1 \\ c &= \theta^2(4a^2t - 4at + 1) - \theta(4at - 4t + 2) + 1 \end{aligned}$$

which leads to the solutions:

$$v_1 = \frac{2t(a\theta - 1)^2}{b + \sqrt{c}}, \quad v_2 = \frac{2t(a\theta - 1)^2}{b - \sqrt{c}}.$$

To select the correct solution, note that $v \geq 0$ by definition. However, if, for example, $u = 0$, (which means $a = 1$), $t = 0.5$ and $\theta = 0.5$, then $v_2 = -1$, whereas v_1 is non-negative for all values of a , θ , and t . Hence, the correct solution for v is:

$$v = \frac{2t(a\theta - 1)^2}{b + \sqrt{c}}.$$

From this, the algorithm for generating pseudorandom numbers from an AMH distribution can be expressed as follows, as in Nelsen [2006]:

1. Generate two independent *Uniform*(0, 1) observations u and t .
2. Set:

$$\begin{aligned} a &= 1 - u \\ b &= -\theta(2at + 1) + 2a^2\theta^2t + 1 \\ c &= \theta^2(4a^2t - 4at + 1) - \theta(4at - 4t + 2) + 1. \end{aligned} \tag{4.1}$$

3. Set

$$v = \frac{2t(a\theta - 1)^2}{b + \sqrt{c}}.$$

4. The desired sample is (u, v) .

Conveniently, the algorithm for generating pseudorandom numbers from a parent copula can be readily extended to generation from a mixture model. This is because the mixture copula is essentially a hierarchical model, as discussed above in Section 2.6. Thus, the *Beta*-mixture copula $C'_\alpha(u, v; b)$ is the bivariate marginal distribution of (u, v) in the following hierarchical distribution:

$$C_\Theta(u, v|\theta) = \frac{uv}{1 - (1 - u)(1 - v)}.$$

where $\Theta = 2X - 1$ and $X \sim \text{Beta}(\alpha, b)$.

Hence, given that the mixing distribution is relatively easy to sample from, the above algorithm can be adapted by adding one step for the generation of the (now random) parameter θ , as follows:

1. Generate two independent $\text{Uniform}(0, 1)$ observations u and t .
2. Generate an independent variable x with distribution $\text{Beta}(\alpha, b)$; set

$$\theta = 2x - 1. \quad (4.2)$$

3. Generate the desired pair via algorithm (4.1)

In this experiment, data pairs are pseudo-randomly generated from each of the AMH model (2.8) and the Beta -mixture AMH model with fixed b parameter (3.17) using algorithms (4.1) and (4.2). In each replication, sample size was fixed at $n = 1000$. A total of 200 replications are performed for each parameter value. A wide selection of parameter values are assigned in data generation, in order to cover a large portion of the dependence coverage of the two copulas.

4.3 Estimation

Once pseudorandom data are generated using algorithms (4.1) and (4.2), the next step is to estimate, using these data, models given by the AMH family (2.8) and the Beta -mixture family with fixed β (3.17). As noted above, in the case of uniform margins, the copula is equivalent to the joint cdf of the distribution. The models can thus be estimated by maximum likelihood. Note that, in both cases there is one parameter to be estimated — θ in the case of the AMH family, and α in the case of the Beta -mixture family.

Thus, for a random sample of size n on (U, V) , the log-likelihood for θ in the AMH model is:

$$\begin{aligned} \log L &= \sum_{i=1}^n \log \frac{\partial^2}{\partial u \partial v} C_{\theta}(u_i, v_i) \\ &= \sum_{i=1}^n \log \frac{\partial^2}{\partial u_i \partial v_i} \left[\frac{u_i v_i}{1 - \theta(1 - u_i)(1 - v_i)} \right] \\ &= \sum_{i=1}^n \log \frac{(1 - u_i)(1 - v_i)\theta^2 + [(1 + u_i)(1 + v_i) - 3]\theta + 1}{[1 - (1 - u_i)(1 - v_i)\theta]^3}. \end{aligned} \quad (4.3)$$

Similarly, for a random sample of size n on (U, V) , the log-likelihood for α in the *Beta*-mixture AMH is:

$$\begin{aligned}
\log L' &= \sum_{i=1}^n \log \frac{\partial^2}{\partial u \partial v} C'_\alpha(u_i, v_i; b) \\
&= \sum_{i=1}^n \log \frac{\partial^2}{\partial u \partial v} \left[\frac{u_i v_i}{1 + (1 - u_i)(1 - v_i)} {}_2F_1(1, \alpha; \alpha + b; s) \right] \\
&= \sum_{i=1}^n \log \frac{2[(1 - u_i) + (1 - v_i)] {}_2F_1(1, \alpha; \alpha + b; s)}{[1 + (1 - u_i)(1 - v_i)]^3} \\
&\quad - \frac{2\alpha[(1 - u_i)(1 - v_i) - 1][(1 - u_i)(1 - v_i) - 3] {}_2F_1(2, \alpha + 1; \alpha + b + 1; s)}{(a + b)[1 + (1 - u_i)(1 - v_i)]^4} \\
&\quad + \frac{8\alpha(1 + a)u_i v_i(1 - u_i)(1 - v_i) {}_2F_1(3, \alpha + 2; \alpha + b + 2; s)}{(a + b)(\alpha + b + 1)[1 + (1 - u_i)(1 - v_i)]^5}
\end{aligned} \tag{4.4}$$

where repeated use has been made of the formula

$$\frac{\partial}{\partial t} {}_2F_1(p, q; r; t) = \frac{pq}{r} {}_2F_1(p + 1, q + 1; r + 1; t).$$

4.4 Results and discussion

Table 4.1 reports a selection of simulation results from estimation of the AMH copula (2.8) and the *Beta* $(\alpha, 1)$ -mixture of the AMH copula, obtained by substitution $b = 1$ into (3.17) to give:

$$C_\alpha(u, v; 1) = \frac{uv}{1 + (1 - u)(1 - v)} {}_2F_1(1, \alpha; \alpha + 1; s).$$

Note that the *Beta* $(\alpha, 1)$ distribution has pdf given by, from (3.1),

$$f(x; \alpha) = \frac{1}{a} x^{\alpha-1}, \quad 0 < x < 1, \quad \alpha > 0,$$

which corresponds to a Power distribution with parameter α . The sample size is fixed at 1000 for each replication, and 200 replications are performed for each parameter value, with sample average results reported.

The first column in the table indicates the true distribution and parameter value used in simulation. The first set (labelled $\alpha = 0.05, \dots, 50$) correspond to when the *Beta* $(\alpha, 1)$ -mixture AMH is the true data-generating process, while the second set (labelled $\theta = -0.9, \dots, 0.9$) are when the AMH is the true data-generating process. The remainder of the table gives the parameter estimates, standard error of estimates, and the estimated likelihoods for each of the two models.

In comparing the fit of the models, it should be noted that a great variety of techniques exist for model selection among copulas. Joe [1997, Section 10.3], for

Table 4.1: Estimates of the AMH model and Beta-mixture AMH model, $b = 1$ from generated data.

Parameter	ρ	AMH		$Beta(\alpha, 1)$ -AMH	
		$\hat{\theta}$	$\log L$	\hat{a}	$\log L'$
$\alpha = 0.05$	-0.2443	-0.8926 (0.0882)	33.4841	0.0530 (0.0444)	33.5225
0.1	-0.2197	-0.8016 (0.1079)	26.9290	0.1012 (0.0581)	27.0240
0.3	-0.1382	-0.5020 (0.1133)	11.4186	0.2927 (0.0857)	11.8461
0.5	-0.0760	-0.2466 (0.1076)	3.3771	0.5198 (0.1174)	4.1080
0.7	-0.0266	-0.0935 (0.1023)	0.9870	0.7053 (0.1448)	1.9480
0.8275	0.0000	-0.0079 (0.0983)	0.5398	0.8324 (0.1637)	1.7140
1	0.0313	0.0938 (0.0929)	1.0275	1.0108 (0.1914)	2.2251
1.5	0.1006	0.3050 (0.0808)	6.3645	1.5532 (0.2831)	7.8602
2	0.1498	0.4207 (0.0727)	12.6902	2.0299 (0.3722)	13.9918
3	0.2157	0.5847 (0.0593)	27.4526	3.1340 (0.5984)	28.4526
4	0.2583	0.6619 (0.0522)	38.2372	4.0300 (0.7958)	39.3186
5	0.2884	0.7232 (0.0462)	49.0132	5.1059 (1.0601)	49.8918
10	0.3639	0.8482 (0.0321)	81.8613	10.1983 (2.5647)	82.4709
20	0.4135	0.9221 (0.0217)	114.5770	22.0767 (7.4427)	114.9090
50	0.4498	0.9674 (0.0130)	150.0340	62.2490 (23.8049)	150.0922
$\theta = -0.9$	-0.2483	-0.8968 (0.0868)	34.1796	0.0503 (0.0432)	34.1919
-0.8	-0.2248	-0.8017 (0.1076)	27.4037	0.0970 (0.0572)	27.3536
-0.5	-0.1489	-0.5010 (0.1103)	11.8508	0.2813 (0.0836)	11.3550
-0.2	-0.0636	-0.1941 (0.1026)	2.3519	0.5625 (0.1241)	1.3157
0	0	0.0008 (0.0947)	0.5757	0.8353 (0.1662)	-0.5523
0.2	0.0703	0.2003 (0.0840)	3.0780	1.2472 (0.2339)	1.7134
0.3	0.1084	0.3019 (0.0780)	6.6382	1.5448 (0.2871)	5.1543
0.5	0.1924	0.4860 (0.0649)	18.7439	2.4063 (0.4535)	17.4131
0.8	0.3451	0.7967 (0.0368)	70.2357	7.5619 (1.7771)	69.3977
0.9	0.4070	0.8982 (0.0245)	106.6634	16.2952 (4.9038)	106.2233

- Notes: (i) Sample size $n = 1000$
(ii) Standard errors in braces
(iii) Estimates are averaged over 200 replications
(iv) Figures to 4dp

example, suggests using the Akaike information criterion (AIC), which evaluates the fit of a model using its log-likelihood, penalised for the number of parameters in the model. A number of other goodness-of-fit criteria, such as the Bayesian information criterion (BIC; also known as the Schwartz information criterion or SIC), are also based on the log-likelihood penalised for the number of parameters in the model. The AIC and BIC measures are given as:

$$AIC = -2 \ln L + 2k \quad (4.5)$$

and

$$BIC = -2 \ln L + k \ln(n) \quad (4.6)$$

where $\ln L$ is the estimated log-likelihood; k is the number of parameters estimated; and n is the sample size. However, note that both the AMH copula and the mixture AMH copula have a single parameter — θ and α respectively. Hence, a comparison of their AIC or BIC statistics is equivalent to a comparison between their log-likelihoods.

Comparing the fits alone, as measured by the maximised log-likelihood values, leads to two observations.

4.4.1 Comparing mixture-copula data with parent copula data

Upon examining the performance of the two models, it can be seen that neither copula model completely dominates the other. Instead, the performance of the true data generating process is always superior to that of the misspecified model, as judged through maximised log-likelihood. For data simulated from the *Beta*-mixture AMH copula, it generally provides a better fit than the parent, AMH copula. Likewise, for data simulated from the AMH copula, the AMH copula generally provides a better fit than the *Beta*-mixture. That each model outperforms the other when it is the true model is an indication that the mixture copula is *not* a generalisation of the parent copula.

This is an important reminder that, contrary to perception, the added flexibility of mixture models arises from extension of the parameter space, *not* from the hierarchical nature of the model. That is, while the mixture model can be viewed as having a parameter that varies, whereas the parent model has the parameter fixed, this alone does not give flexibility to the model. The mixture model is, generally, a separate model with a different functional form and which generally does not nest the parent model. Flexibility is added only if parameter-mixing extends the parameter space, as in the *Beta-Binomial* example.

In the present case, the *Beta*($\alpha, 1$)-mixture AMH copula is restricted to a single dependence parameter to ensure identification. As a result, the parameter space that

indexes both the mixture and the parent families is of the same dimension. Thus, the mixture copula is not a generalisation of the parent copula. Both models are competing on an equal footing in the parametrically non-nested sense, as shown by the estimation results.

The implication of this result is that parameter-mixing does not yield a *general* modelling advantage. Rather, the mixture copula will out-perform the parent copula only where the data displays a dependence structure better captured by the mixture copula.

4.4.2 Comparing results at different dependence levels

As found above in Chapter 3, at the limits of their dependence coverage, the *Beta*-mixture AMH family and the parent AMH family converge. Both competing families of copulas have the same extremals, $C_{-1}(u, v)$ and $C_{+1}(u, v)$. As we approach the extremes of the dependence coverage, therefore, the two copulas should become increasingly difficult to differentiate. It is thus interesting to examine whether the dependence level of the simulated data has an impact on any difference in performance between the two models.

From Table 4.1, we can see that any differences in fit as measured by the maximised log-likelihoods vanish as the true model approaches the limits of the range of dependence coverage. When the mixture copula is the true model, differences in the maximised log-likelihoods disappear as α tends to either extreme of the parameter space. Likewise, when the AMH copula is the true model, differences in the maximised log-likelihoods disappear as $\theta \rightarrow \pm 1$. As expected, as the parameter approaches its extreme value, the copulas become increasingly difficult to distinguish, even though one of them is always misspecified.

A related observation is that the greatest difference in fit occurs near the centre of the dependence range. From the table above, the fit of the true model exceeds the fit of the misspecified model by the greatest amount when $\alpha = 1.5$ under the *Beta*($\alpha, 1$)-mixture AMH, and when $\theta = 0.3$ under the AMH, which correspond to a Spearman's rho measure of 0.1006 and 0.1084 respectively. The dependence coverage of both models is $[-0.2711, 0.4784]$, with midpoint 0.1037. That the greatest advantage appears near the centre of the dependence coverage suggests that should any advantage be derived by applying parameter-mixing to dependence parameters, then that will require the data to “cooperate” by exhibiting sample dependence in the “middle” of the range of dependence coverage of the parent family of copulas.

In summary, the simulation results indicate two things. Firstly, the mixture copula and the parent copula are parametrically non-nested. Each performs better when it

is (or is closer to) the true distribution. Secondly, modelling advantage is greatest when data exhibits dependence in the “middle” of the range of dependence coverage of the parent family of copulas; the advantage diminishes as dependence approaches the limits of the range of dependence coverage.