End-term project: Advanced Statistical Computing 2020

Deadline October 27th 2020

1 Introduction

In statistics, correlation or dependence refers to any statistical association, to which a pair of random variables are linearly related. When dependence is not modelled, resulting statistical inference may be misleading. This can be costly, especially in the financial or medical sector.

Copula models have become increasingly attractive due to their aptitude to represent any multivariate joint distribution in terms of univariate marginal distribution functions and a copula which describes the dependence structure between the random variables. That is, the joint distribution of a random vector can be written in terms of marginal distribution functions and a copula function. The marginal describe the behaviour of the random variables and the copula function the dependency structure between the random variables. As such, the main purpose of copulas is to describe the interrelation of several random variables and represent the joint distribution as a composition of marginals and a copula. This makes copula models an attractive tool when more information about marginal behaviour is available instead of the joint, common in finance and insurance.

In this paper the potential of a copula model in the setting of risk optimization of an insurance company (AVN) is examined. AVN, is an insurance company which sells two insurance policies (also called business lines) to corporate clients:

- Professional Liability Insurance (PLI)
- Workers' Compensation (WC)

Over the course of the last two years, these two business lines were simultaneously affected by huge claims related to a single client. The client involved held policies from both business lines, which led to significant claims. In order to protect themselves from future risks related to huge insurance claims, AVN approaches a reinsurance company (RC) with the following idea:
For some threshold $t = 100, 110, ..., 200$:

- If $PLI + WC \leq t$, ANV pays the claim themselves
- If $PLI + WC > t$, the RC pays the claim

For the same threshold $t = 100, 110, ..., 200$ the insurance company request the following price $P(t)$ in million euros:

$$P(t) = 40000 \times \exp\left(\frac{-t}{7}\right)$$

, which is a decreasing function of t as seen in Figure 1.

Figure 1: Price (in million euros) as a function of threshold t as requested by the Reinsurance Company (RC).

If the expected reinsurance policy value V(t) exceeds the requested Price P(t), AVN will purchase the policy. That is if:

$$V(t) = E[(PLI + WC)1(PLI + WC > t)] > P(t) \tag{1}$$

In order to determine whether a reinsurance policy should be purchased to transfer a portions of AVN's own risk portfolios to the reinsurance company across different thresholds (t), a copula model is used, where the expected reinsurance payout value $V(t)$ is approximated using a simulation study. In more detail, in this paper we first verify whether the copula estimation method is correct. Then we conduct a simulation to develop an understanding of the inner workings of parameter estimation. After this, we approximate $V(t)$ using an Monte Carlo (MC) simulation and Importance Sampling (IS). Lastly, a Bootstrap procedure is conducted to compute a 80% Confidence Intervals (CIs) across all threshold values t to ensure that the advice given to AVN is correct.

2 Methodology

2.1 Data

The dataset contains data from clients that occurred losses in both PLI (X_1 and WC (X_2 business lines. There are three columns:

- ID: client ID
- PLI: loss incurred for PLI (in million euros)
- WC: loss incurred for WC (in million euros)

and a total of 648 observations (N = 648) observations. The variables PLI and WC have a linearly positive relationship (Figure 2) and a Pearson correlation coefficient of $\rho = 0.528$.

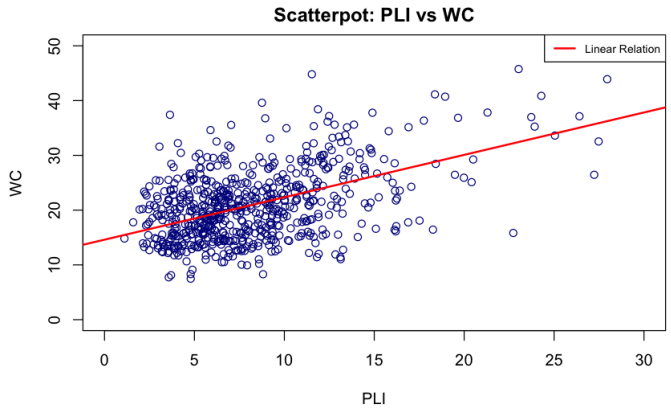


Figure 2: Scatter plot of the two variables PLI and WC, including a regression line.

2.2 The Model

In order to approximate $V(t)$, using a copula model the joint distribution of the two businesses F_{X_1,X_2} is decomposed into the joint density using the copula model as followed :

$$f_{X_1,X_2}(x_1,x_2) = f_{X_1}(x_1)\, f_{X_2}(x_2)\, c\left(F_{X_1}(x_1), F_{X_2}(x_2)\right)$$

, where $c(u_1,u_2)$ is the copula density which induces the dependencies between the marginal densities of f_{X_1} and f_{X_2}. The function c is the joint density of the probability integral transformation of $U_1 = F_{X_1}(X_1)$ and $U_2 = F_{X_2}(X_2)$. Visibly, the copula models the dependence between the random variables. This is because a copula is a multivariate cumulative distribution function, which represents the marginal probability distribution in terms of a uniformly distributed univariate variables on the interval $[0,1]$. As such a copula decomposes a joint density into marginals, which are uncorrelated to each other and binds them together using a copula function, which also specifies the dependence structure.

Preliminary experiments on the data suggest the following attributes and parametric model for the marginals f_{X_1}, f_{X_2} and the copula c:

- $f_{X_1}(\cdot;\mu_1,\sigma_1) \sim \text{Lognormal}(\mu_1,\sigma_1), \mu_1 \in \mathbb{R}, \sigma_1 > 0$
- $f_{X_2}(\cdot;\mu_2,\sigma_2) \sim \text{Lognormal}(\mu_2,\sigma_2), \mu_2 \in \mathbb{R}, \sigma_2 > 0$
- $c(\cdot;\theta) \sim \text{Joe}(\theta), \theta \geq 1$

2.3 Analysis

Maximum likelihood estimation (MLE) is used to estimate the unknown parameters of marginals f_{X_j} for $j = 1,2$ and the unknown Joe copula parameter θ. As for the marginals MLE is appropriate since we can minimize the negative log-likelihood. For the copula parameter θ MLE is appropriate, because it is possible to derive log-likelihood function from the copula density over which we can optimize θ. After obtaining all unknown parameters, a simulation study is conducted to examine the inner workings of the estimation procedure and to verify whether estimations are plausible. $V(t)$ is then approximated using a MC simulation and IS. Lastly a bootstrapping procedure is used to compute the 80% CIs to ascertain whether a reinsurance policy should be purchased.
Since ML estimates are needed for the simulation study and describe attributes of the data i.e the marginals, these are included in the method section instead of the results.

2.3.1 Maximum Likelihood Estimation

For both marginals f_{Xj} for $j = 1, 2$ the unknown μ_j and σ_j are estimated by minimizing the negative log-likelihood. Since the marginals have a log-normal distribution, the MLE equations for the log-normal distribution are:

$$\hat{\mu_n} = \frac{1}{n}\sum_{j=1}^{n} \ln x_j, \quad \hat{\sigma}_n^2 = \frac{1}{n}\sum_{j=1}^{n} (\ln x_j - \hat{\mu})^2$$

The mean and standard deviation of X_1 and X_1 are initialized as starting values for the optimization procedure using the R base function optim(). The starting values and resulting maximum likelihood estimates for the unknown parameters of the marginals are illustrated in Table 1.

	μ	σ
f_{X_1}	8.278	4.641
f_{X_2}	21.019	6.812

	$\hat{\mu}$	$\hat{\sigma}$
f_{X_1}	1.982	0.513
f_{X_2}	2.997	0.311

Table 1: Starting values of the MLE optimization for the unknown parameters of the marginals (left) and resulting ML estimates of parameters (right).

The unknown parameter θ of the Joe copula is also estimated using MLE. However, since the probability integral transform $U_j = F_{X_j}(X_j)$ for $j = 1, 2$ is not observed, pseudo observations $\hat{U}_j = F_{\hat{\mu}_i, \hat{\sigma}_j}(X_j), j = 1, 2$ are generated and used instead. This similarly works, since we can derive a log-likelihood function from the copula density over which we can then optimize the unknown Joe copula parameter θ for $\hat{U}_j$. Using the maximum likelihood estimates for the marginal, results in an MLE for the unknown Joe copula parameter $\theta = 1.608$, which will be used for further analysis and the simulation study. To estimate θ, the optim() function was used as well.
In the following sections, we will simply refer to the ML estimates as determined above as $\hat{\mu_j}, \hat{\sigma_j}$ and $\hat{\theta_j}$.

2.4 Monte Carlo Simulation

Monte Carlo (MC) methods use simulations to study properties of a random mechanism and exploit the law of large of large numbers to obtain a reasonable estimate of the expected value. That is, the more samples we generate, the closer the estimated value is to the true expectation. The MC algorithm can be summarized using the following steps:

Algorithm 1: Monte Carlo Simulation

Given that we wish to compute $\mathrm{E}[g(\boldsymbol{X})]$ with $\boldsymbol{X} \sim F$ and $\boldsymbol{X} \in \mathbb{R}^d$

1. Simulate Data $\boldsymbol{x}_1, \ldots, \boldsymbol{X}_N \overset{iid}{\sim} F$
2. Compute the average $M_n = \frac{1}{N}\sum_{i=1}^{N} g(X_i)$

Law of larger numbers $M_N \to_p \mathrm{E}[g(\boldsymbol{X})]$

That is using an MC simulation we can approximate the expected value of $V(t)$ via the following proposal function $g(x)$:

$$V(t) = \mathbb{E}[g(x)] = \mathbb{E}\left[1\left(X_1 + X_2 > t\right)\left(X_1 + X_1\right)\right], \quad t \in \mathbb{R}$$

This means for each MC simulation the values of the MC simulation, where $(X1 + X2 > t)$ are summed, all other values are set to 0 and then divided by the sample size N. This way,

it is possible to derive an estimate for $V(t)$. In this MC simulation $N = 10^5$ MC samples are drawn.

2.5 Importance Sampling

MC simulation may be inefficient when only a few samples effect M_N. That is some samples have a higher impact than others. However, Importance Sampling (IS) is a method that overcomes this problem, by sampling from another distribution and then computing the weighted average. That way the importance or influence of samples that have an impact is increased and the approximation error reduced. The IS algorithm can be summarized using the following steps:

Algorithm 2: Importance Sampling

Given that we wish to compute $\mathrm{E}[g(\boldsymbol{X})]$ with $\boldsymbol{X} \sim f_{\boldsymbol{X}}$
Sample from another distribution $\boldsymbol{Y} \sim f_{\boldsymbol{Y}}$

1. Sample $Y_1, \ldots, Y_n$ from $f_{\boldsymbol{X}}$
2. Compute the average $M_N = \frac{1}{N} \sum_{i=1}^{N} g\left(\boldsymbol{Y}_i\right) \frac{f_X(\boldsymbol{Y}_i)}{f_Y(\boldsymbol{Y}_i)}$

$$\text{So } \mathrm{E}[g(\boldsymbol{X})] = \mathrm{E}\left[g(\boldsymbol{Y}) \frac{f_X(\boldsymbol{Y})}{f_Y(\boldsymbol{Y})}\right]$$

The density $f_{\boldsymbol{Y}}$ is chosen to be close to $g(\boldsymbol{y}) f_{\boldsymbol{X}}(\boldsymbol{y})$. Data for the simulation is generate from the joint model with parameters $\mu_1 = \hat{\mu_1} + 2, \mu_2 = \hat{\mu_1} + 2, \sigma_1 = \hat{\sigma_1}, \sigma_2 = \hat{\sigma_2}, \theta = \hat{\theta}$ to ensure that that the event of interest becomes more likely.

2.6 Bootstrap

In order to assign a measure of accuracy to the estimated $V(t)$ a 80% confidence interval (CI) is computed. The number of bootstrap samples is set to $B = 1000$. For each bootstrap sample observations are generated using the joint with $\mu_1 = \hat{\mu_1} + 2, \mu_2 = \hat{\mu_1} + 2, \sigma_1 = \hat{\sigma_1}, \sigma_2 = \hat{\sigma_2}, \theta = \hat{\theta}$. The Bootstrap algorithm can be described using the following steps:

Algorithm 3: Boostrap

1. Create bootstrap indices: Randomly sample (with replacement) values from 1 : N (Total number of observations)
2. Select the observations according to bootstrap indices and compute MLE estimates of the marginals: $\mu_1 = \hat{\mu_1}, \mu_2 = \hat{\mu_1}, \sigma_1 = \hat{\sigma_1}, \sigma_2 = \hat{\sigma_2}, \theta = \hat{\theta}$
3. Generate observations from the joint with parameters $\mu_1 = \hat{\mu_1} + 2, \mu_2 = \hat{\mu_1} + 2, \sigma_1 = \hat{\sigma_1}, \sigma_2 = \hat{\sigma_2}, \theta = \hat{\theta}$
4. 6. Compute the expected value of V (t) using Importance Sampling across each threshold t.
5. Repeat B times

The resulting matrix has B number of rows and t number of columns for which the lower and upper bound of the 80% is computed by sorting values and finding the cutoff at 10% and 90%.

3 Simulation Study

3.1 Examining the Estimated Model

To verify whether model estimations are plausible, data was simulated from an estimated model. If the model is plausible, generated data and observed data should be alike. A comparison between observed and simulated values can be see in the Scatter plot Figure 3. Visibly, using ML estimates as input values ($\mu_1 = \hat{\mu_1}, \mu_2 = \hat{\mu_1}, \sigma_1 = \hat{\sigma_1}, \sigma_2 = \hat{\sigma_2}, \theta = \hat{\theta}$) produces simulated data almost identical to the observed data.

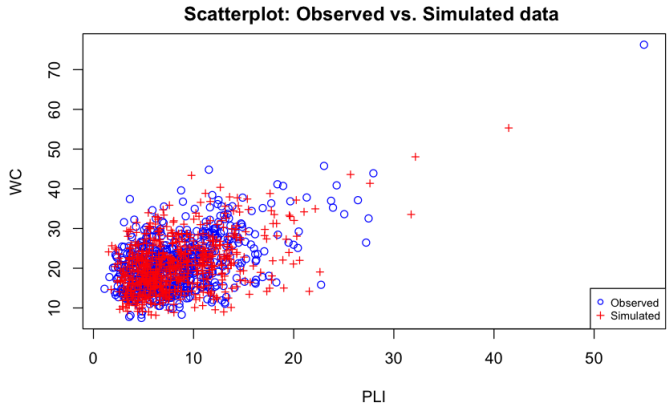
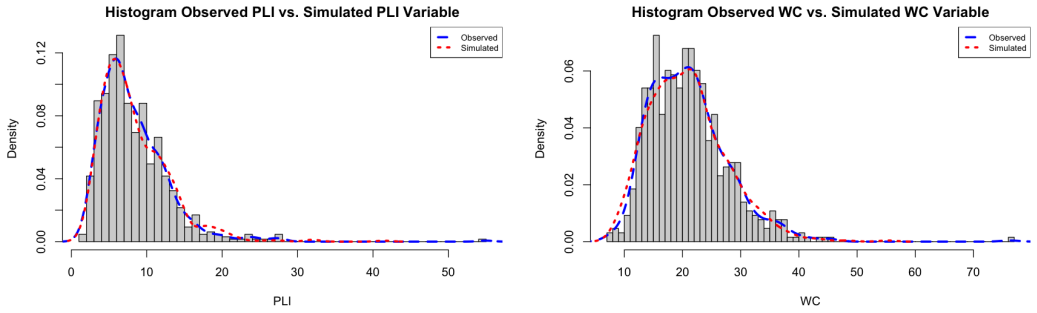


Figure 3: Scatter plot of observed vs. simulated values.

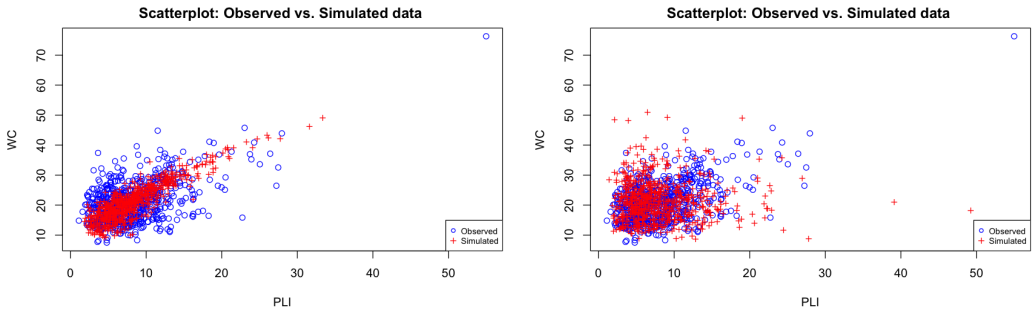
The almost identical similarity is further supported by examining the histograms in Figure 4. The simulated density as shown in red is almost identical to the observed density in blue.



(a) X_1 (b) X_2

Figure 4: Histogram of observed X_1 in A and X_1 in B, depicting both the density of observed values and simulated values

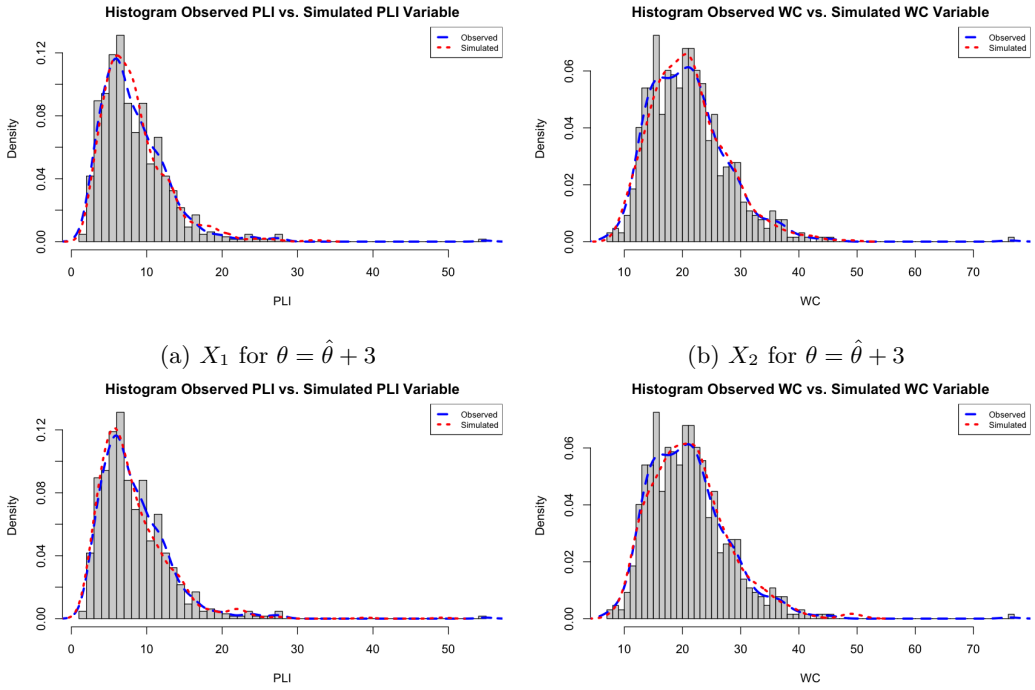
To ensure that the estimated model is correct, properties of the simulated data are examined across different input values. Since θ is the copula parameter, a change in θ should result in a changes in the dependence structure of the simulated data. To test this, data from two estimated models were generated. One model with a large value of θ and one with a small value. The resulting change in dependence from the first model with $\mu_1 = \hat{\mu_1}, \mu_2 = \hat{\mu_1}, \sigma_1 = \hat{\sigma_1}, \sigma_2 = \hat{\sigma_2}, \theta = \hat{\theta}+3$ can be examined in Figure 5. Visibly, the simulated data points cluster together, indicating a higher correlation structure as compared to the previous example in Figure 3 and Figure 4.



(a) Model with $\theta = \hat{\theta} + 3$

(b) Model with $\theta = 1$

Figure 5: Scatter plot of observed vs. simulated data with a relatively large copula parameter $\theta = \hat{\theta} + 3$ in A and a small copula parameter $\theta = 1$



(a) X_1 for $\theta = \hat{\theta} + 3$

(b) X_2 for $\theta = \hat{\theta} + 3$

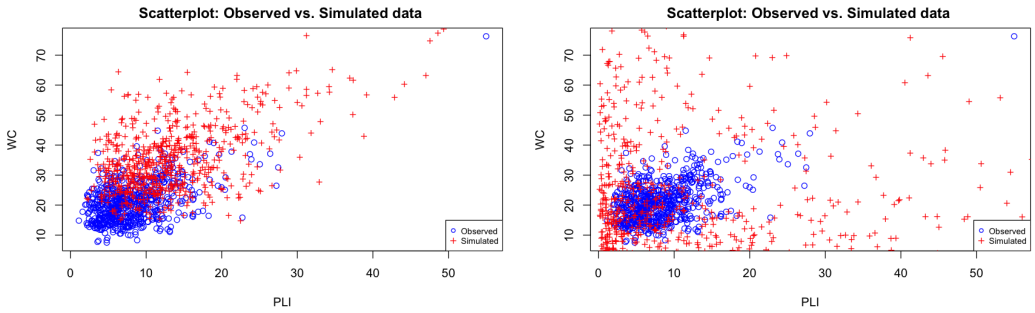
(c) X_1 for $\theta = 1$

(d) X_2 for $\theta = 1$

Figure 6: Histogram of observed $X1$ and $X2$, depicting both the density of observed values and simulated value of $\theta = \hat{\theta} + 3$ from (a) to (b) and $\theta = 1$ from (c) to (d)

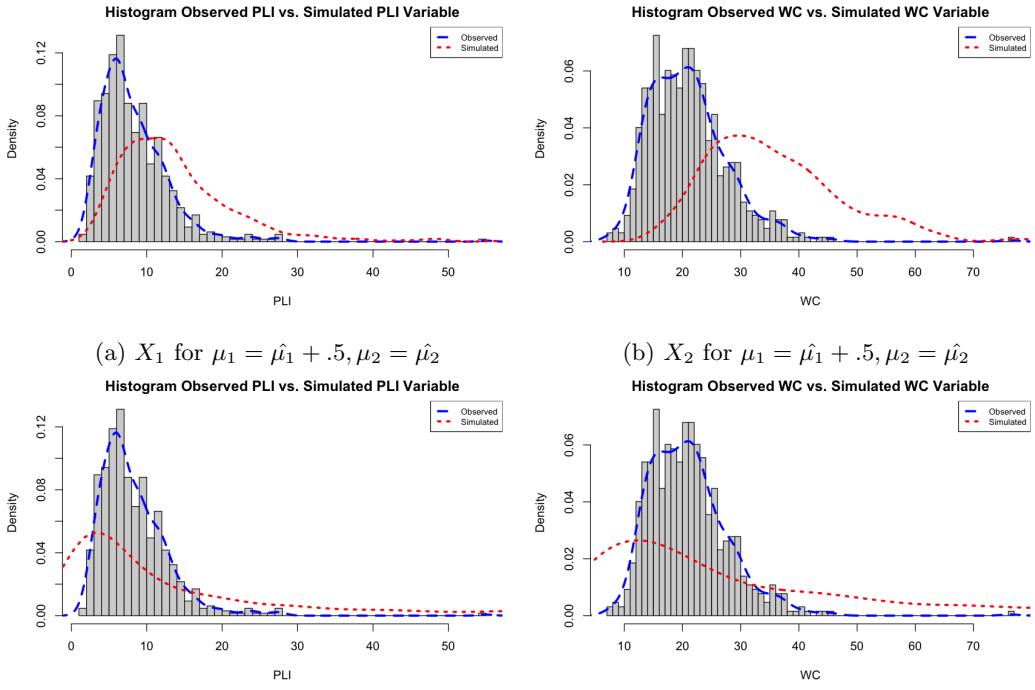
Densities are not effected, which is to be expected, since marginals are not influenced by the copula parameter θ as shown in Figure 6. To support this, simulations from two more models are generated. One model to illustrate the influence of the mean another for the variance. Simulations results are graphically depicted in Figure 7 with respective densities in Figure 8. A change in the mean changes the location of simulated values as shown in Figure 7 (a) and the variance the spread visible in Figure 7 (b).

As can be seen in Figure 8, a change in mean and variance results in visible changes in the density. The location of the density changes with a change in the mean and as the variance increases the distribution becomes more spread and flat. This is to expected as mean and variance control the location and spread.



(a) Model with $\mu_1 = \hat{\mu_1} + .5, \mu_2 = \hat{\mu_2}$ (b) Model with $\sigma_1 = \hat{\sigma_1} \times 3, \sigma_2 = \hat{\sigma_2} \times 3$

Figure 7: Scatter plot of observed vs. simulated data with changes in μ and σ.



(a) X_1 for $\mu_1 = \hat{\mu_1} + .5, \mu_2 = \hat{\mu_2}$ (b) X_2 for $\mu_1 = \hat{\mu_1} + .5, \mu_2 = \hat{\mu_2}$

(c) X_1 for $\sigma_1 = \hat{\sigma_1} \times 3, \sigma_2 = \hat{\sigma_2} \times 3$ (d) X_2 for $\sigma_1 = \hat{\sigma_1} \times 3, \sigma_2 = \hat{\sigma_2} \times 3$

Figure 8: Histogram of observed $X1$ and $X2$, depicting both the density of observed values and simulated value of $\mu_1 = \hat{\mu_1} + .5, \mu_2 = \hat{\mu_1} + .5$ from (a) to (b) and $\sigma_1 = \hat{\sigma_1} \times 3, \sigma_2 = \hat{\sigma_1} \times 3$ from (c) to (d).

3.2 Examining Method of Estimation

After having verified that the estimated model is correct a simulation study is conducted to better understand the inner workings of the method of parameter estimation. For the simulation study the parameters were fixed to $\mu_1 = 1, \mu_2 = 3, \sigma_1 = 2, \sigma_2 = 0.5, \theta = 2$ and for r = 1,...,100 repetitions:

(i) n observations $(X_{i,1}, X_{i,2})$, $i = 1, ..., n$ are simulated from from the joint model.

(ii) and the model parameters are fit while tracking the time it takes using the system.time() function in R

The whole procedure was repeated for varying numbers of $n = 200, 500, 1000$. Results for

Root Mean Square Error (RMSE) and average computation time are depicted in Figure 9 and Figure 10 respectively.

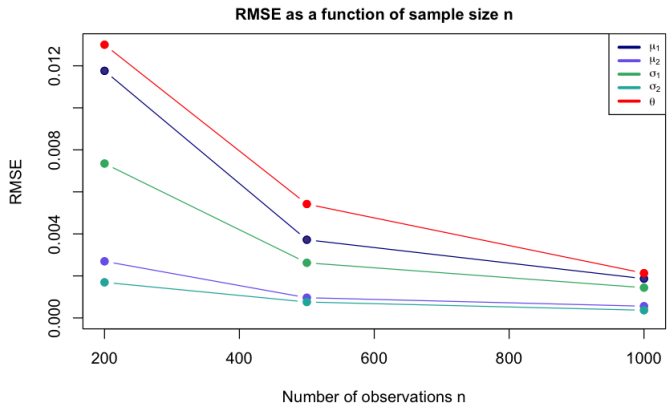


Figure 9: RMSE as a function of observations n for all unknown parameters of the copula model.

Figure 9 shows that across all parameters the RMSE is a decreasing function of n. As observations n increases, so does the certainty and accuracy of the model estimates. For $n = 200$, parameters μ_1 and θ are the hardest parameters to estimate as indicated by their relatively high RMSE. When examining the starting values for μ_1 and σ_1 and comparing them with μ_2 and σ_2, this is to be expected, since the starting values for $\sigma_1 = 2$ is larger than $\sigma_2 = .5$, which leads to larger RMSE differences in the estimates of μ_1 and μ_2. If $\sigma_1 = \sigma_2 = .5$, similar estimates across n should be observed (See Appendix Figure 14, which is part an auxiliary simulation to support this argument not part of this specific simulation). As for θ, the high RMSE for $n = 200$ can be explained due to the fact that this parameter is hard to estimate, especially when sample size is small. As sample size increases, a marked increase in accuracy (lower RMSE) is observed, which supports this argument. (See Appendix Figure 15 for an auxiliary simulation to further supportthis argument not part of this specific simulation).

In Figure 10 the average computing time over n is depicted. Computation time is almost linearly increasing as sample size n increases. However, at $n = 500$ the slope decreases, indicating that the average computational time decrease as n becomes sufficiently large.

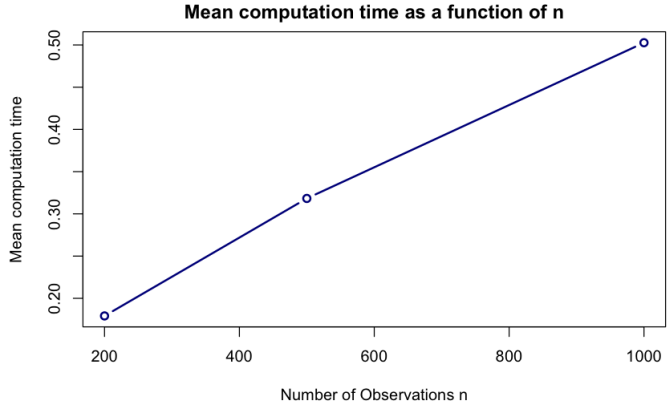


Figure 10: Average computation time as a function of observations n.

Results for the simulations are summarized in Table 2 below. For $n = 200, 500, 1000$ the

average computation time was 0.1046, 0.1986, 0.3193 respectively. The slope at $n = 500$ is equal to 0.000313 and at $n = 1000$ is equal to 0.0002415, which shows that the slope decreases as n becomes sufficiently large.

	$n = 200$	$n = 500$	$n = 1000$
Average time	0.1046	0.1986	0.3193
Increase in time	...	0.0940	0.1208
Slope	...	0.000313	0.0002415

Table 2: Table summarizing the results of the simulation study.

4 Results

4.1 Monte Carlo Simulation

As model estimations are correct $V(t)$ is now approximated using a MC simulation. Samples of size $N = 10^5$ are generated from the estimated model using ML estimates from table 1. Results from the MC simulations are shown in Figure 11. Expected reinsurance payout $V(t)$ and price $P(t)$ are graphed. From the graph it follows that $V(t) > P(t)$ for $t = 100, ..., 130$. The optimal values for the threshold t is determined by finding the largest vertical distance between $V(t)$ and $P(t)$ for $t = 100, ..., 130$. Therefore, according to the MC simulation AVN should purchase a reinsurance for $t = 120$. However due to the fact that $V(t)$ values are quite noisy, because we are dealing with highly improbable events, conclusions drawn from the MC simulation may be misleading. As such values are computed again using IS by generating data from a model that makes events more likely.

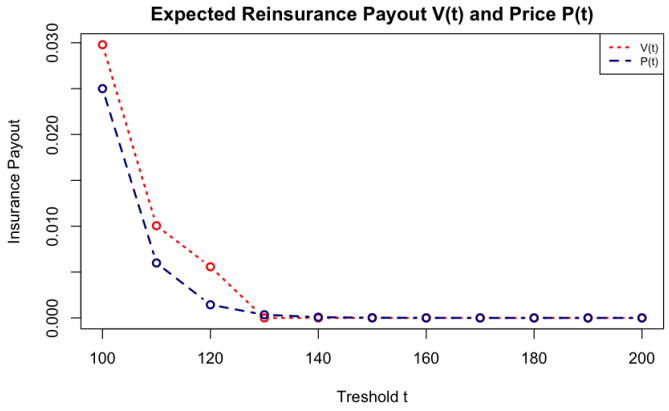


Figure 11: MC Simulation of the expected reinsurance payout $V(t)$ and price $P(t)$ as a function of threshold t.

4.2 Importance Sampling

Values are generated from a model with $\mu_1 = \hat{\mu_1} + 2, \mu_2 = \hat{\mu_2}, \sigma_1 = \hat{\sigma_1}, \sigma_2 = \hat{\sigma_2}, \theta = \hat{\theta}$ in order to make events more probable. Results of the IS procedure can be examined in Figure 12. The largest vertical distance between $V(t)$ and $P(t)$ is at $t = 110$. Therefore, according to the results from the IS procedure AVN should purchase a reinsurance policy for the threshold equal to 110.

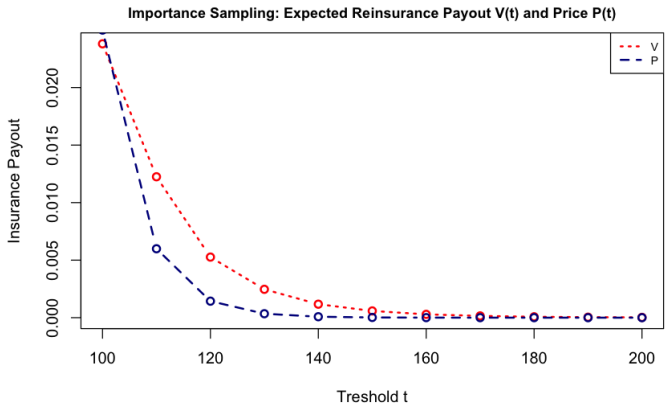
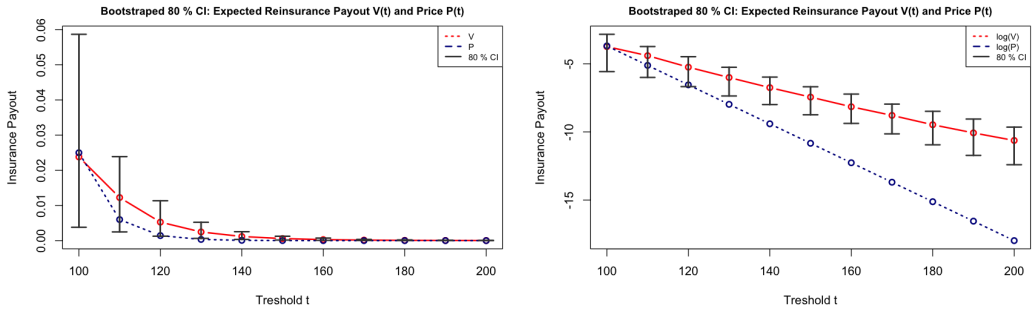


Figure 12: Importance Sampling: Expected value $V(t)$ and price $P(t)$ as a function of t.

4.3 Bootstrap

To ensure that AVN can protect themselves against future large insurance claims 80% CIs are computed for each threshold value t. Results from the Bootstrap procedure and the resulting 80% CIs can be seen in Figure 13. Values for $V(t), P(t)$ and the CIs are depicted both on the original scale and the log scale.



(a) Original scale)

(b) Log scale

Figure 13: Bootstrap 80% CI of the expected value $V(t)$ as a function of t on the original scale (left) and log scaled (right). Price $P(t)$ as a function of t depicted for comparison.

When examining the CIs, we observe that for low values of t there is a lot of uncertainty regarding the $V(t)$ estimate. This can be seen from the wide CIs for low values of t. As such the previous recommendation from the IS procedure to purchase a reinsurance policy at $t = 110$ has to be corrected. On one hand this is because at $t = 110$ the lower bound of the CI intersects with the $P(t)$. This implies that a reinsurance policy may not be profitable for AVN, since expected payout $V(t)$ and price $P(t)$ can overlap. That is AVN and RC may break even or AVN may lose money since the expected payout may be lower than the price for the reinsurance. On the other hand however, AVN has the possibility of substantially making more profits, because the upper bound of the CI is high and far from $P(t)$. Such a decision is however risky and AVN wants to reinsure themselves to transfer a part of their own risk portfolios to the reinsurance company and not increase it. As such this decision is not advised.

Based on these considerations, the optimal value for t is t = 130, since the expected payout $V(t)$ will always be larger than the price for the reinsurance $P(t)$. Furthermore,

this threshold yields the third highest profit obtainable and the lower bound of the CI does not intersect with the $P(t)$ as shown on the log scale in Figure 13b. Therefore, AVN should purchase a reinsurance policy for $t = 130$, because expected payout is larger than the reinsurance price, meaning that AVN will make a profit and successfully transfer their own risk portfolio to the reinsurance company.

Appendix

The results of an auxiliary three additional auxiliary simulations not part of the original simulation study are graphically depicted in Figure 14 and Figure 15. Parameter values were fixed at $\mu_1 = 1, \mu_2 = 3, \sigma_1 = \sigma_2 = 0.5, \theta = 1.6$. As expected if $\sigma_1 = \sigma_2$ similar estimates are observed for μ_1, μ_2, σ_1 and σ_2 as predicted. As for θ, a similar trajectory of RMSE as a function of n is observed. Therefore showing that θ is the hardest parameter to estimate with a similar RMSE trajectory over n regardless of which starting value is used.

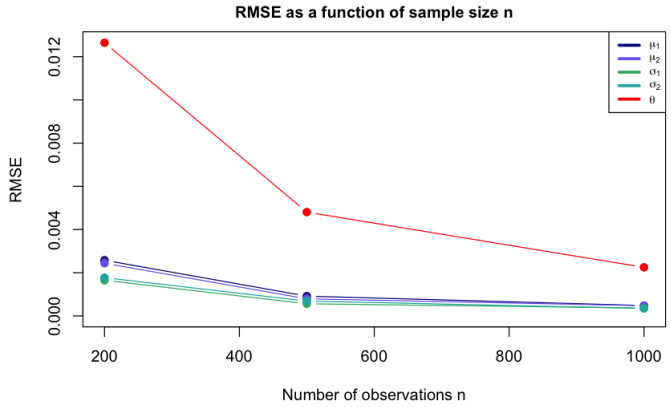
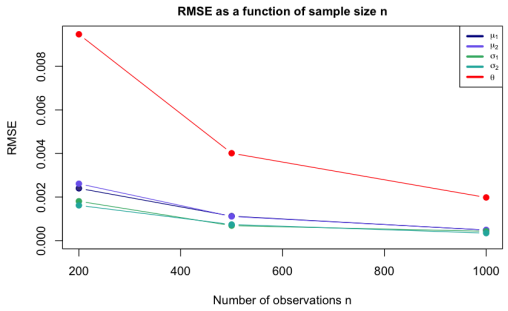
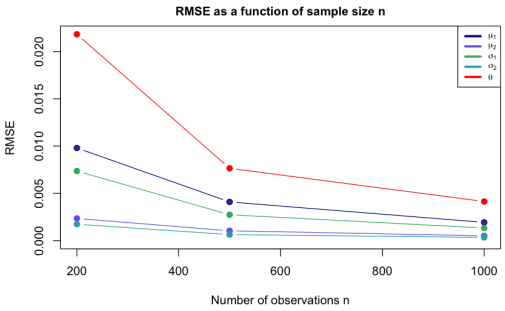


Figure 14: RMSE as a function of observations n with $\mu_1 = 1, \mu_2 = 3, \sigma_1 = \sigma_2 = 0.5, \theta = 2$.



(a) Simulation with $\mu_1 = 1, \mu_2 = 3$, $\sigma_1 = \sigma_2 = 0.5, \theta = 1.6$.)



(b) Simulation with $\mu_1 = 1, \mu_2 = 3$, $\sigma_1 = \sigma_2 = 0.5, \theta = 5$.

Figure 15: RMSE as a function of observations n using two different θ values.