Statistics for Finance and Insurance Assignment 1: Copulas

*Pieter Pijls**

Academic year 2017-2018

Introduction

We investigate a dataset consisting of losses (`loss`) and expenses (`expense`) for n clients of an insurance company. In the first part we will explore the data. We will also create different graphs to analyze the `loss` and `expense` data. In the second part, we will create a model by fitting a Pareto distribution to the variables `loss` and `expense`. In addition, we will discuss the goodness of the fit. In the final part, we will model the dependence structure between the variables. This will be done by fiting a Clayton $C^{\mathrm{Cl}}(u_1, u_2)$ and Gumbel $C^{\mathrm{Cu}}(u_1, u_2)$ copula to the data.

Data exploration

First, we plot the histogram of the `loss` and `expense` variable in Figure 1 to get an idea of the underlying distribution. The histograms illustrate that both variables contain a heavy right tail. We observe extreme outliers which are in many cases censored.

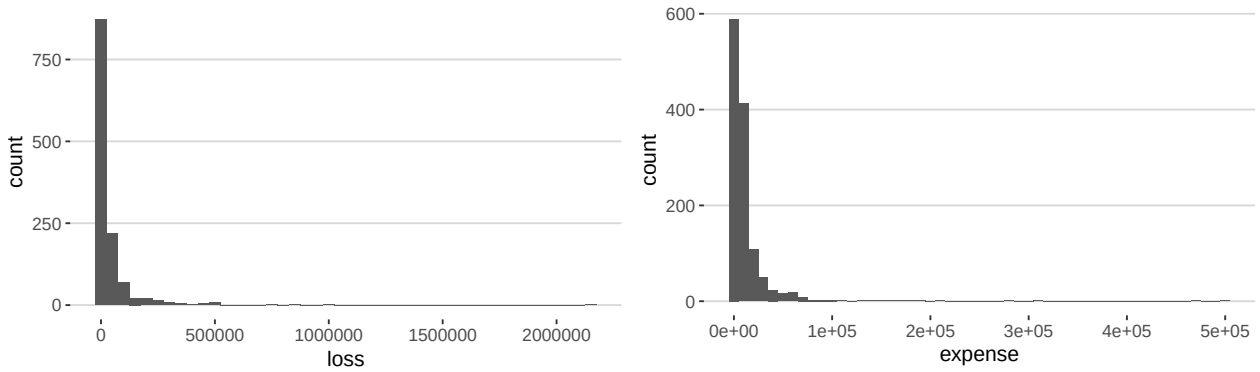

Figure 1: Histogram of the loss and expense variable

In Figure 2 we create a scatterplot where we take the logs the variables `loss` and `expense` to investigate the dependence structure. We take the log transformation to enhance the visualization of the `loss` and `expense` observations. This log transformation preserves the order of the observations while making outliers less extreme. The censored and non-censored obsevations are seperated using different colours. The scatterplot shows a positive dependence structure between the two variables.

*r0387948 (Faculty of Economics and Business, KU Leuven, Leuven, Belgium)

Figure 2: Scatterplot of loss and expense variable

Next, we compute the Pearson correlation coefficient ρ between `loss` and `expense` which is equal to 0.39. The coefficient ρ shows that our observation was correct as there exists a positive linear dependence between the two variables.

In this section we will briefly discuss some features of the observed sample. First, we compute the mean and standard deviation of `loss` and `expense` with the function `sd`. The standard deviations contain high values. Next, we calculate the skewness of `loss` and `expense` with the function `skewness`. We observe a positive skewness which indicates that the right tail is heavier than the left tail. Also the skewness statistic indicates both variables are highly asymmetrically distributed. These statistics show that the variables expense and `loss` have some outliers. Both variables have a very high maximum and a large difference between the median and the mean.

Table 1: Summary Statistics

	Loss	Expense
Mean	4.287112e+04	12804.89120
Standard Deviation	1.097272e+05	29763.68892
Skewness	8.890431e+00	9.17453
Kurtosis	1.301313e+02	118.52390

Modelling the maginals

Here, we fit a Pareto distirbution on each of the marginals by means of maximum likelihood estimation (MLE). First, we create the log-likelihood function for the Pareto model $\mathcal{L}(\alpha, \beta)$ taking censoring into account. The general log-likelihood function $\mathcal{L}$ is denoted as as:

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} (f(x_i)^{1-\delta_i} \cdot (1 - F(c_i)^{\delta_i},$$

where c_i is the observed claim amount in case of censoring (i.e. when $\delta_i = 1$) of the ith policy.

Inserting the cdf F and pdf f of the Pareto distribution with the shape parameter α and scale parameter β gives us:

$$\mathcal{L}(\alpha, \beta) = \prod_{i=1}^{n} \left(\alpha \frac{\beta^\alpha}{x_i^{\alpha+1}} \right)^{1-\delta_i} \cdot \left(\frac{\beta}{c_i} \right)^{\alpha\delta_i}.$$

First, we create the function `loglik.pareto.loss` (see R code) to estimate the parameters of the Pareto distribution using Maximum Likelihood taking censoring into account. Next, we fit a pareto distribution on the variables `loss`. Similarly, we fit a pareto distribution to `expense` using the function `fitdist`. Normally, we use the method of moments for the starting values of α and β. However, another appropriate option is to perform several optimizations and give in the values of previous optimizations as the starting value in the next optimization. After different optimizations the values converge to the optimal values for the parameters α and β. Finally, we take a look at the shape parameter α and scale parameter β of the Pareto models. The shape parameter α of losses is lower than the expenses, which means that the expenses have more observations in the right tail. The scale parameters β are close to each other.

Table 2: Parameters Pareto Model

	Loss	Expense
alpha	1.081348	2.145617
beta	13204.738287	14354.356099

Next, we compare the empirical cdf with the Pareto model for the `loss` variable to check the goodness-of-fit using the estimated values for α and β. From Figure 3 we observe that the model gives us an adequate fit. When we compare the empirical cdf with the Pareto model for the `expense` variable we also have an adequate fit.

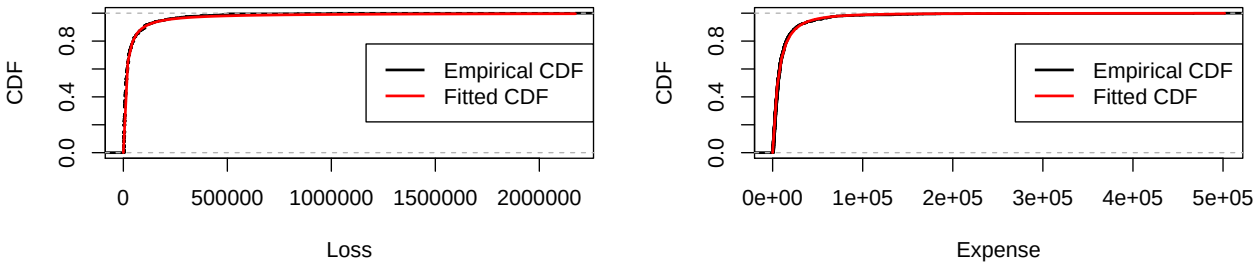

Figure 3: Comparison of Empirical CDF and Pareto CDF

Finally, we compute the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) for both models.

Table 3: Information Criteria

	Loss	Expense
AIC	27569.22	25691.09
BIC	27579.48	25701.35

Modeling the Dependence Structure

Pseudo-observations

Now we fitted the marginal distributions we can model the dependence structure between the variables `loss` and `expense`. We calculate the pseudo-observations `u1` and `u2` using the estimated parameters of our Pareto

model. Notice, censoring is taken into account in `u1` as it is modeled with the parameters of the censored Pareto model. In Figure 4 we construct a histogram of the the pseudo-observations. We observe some deviations from uniform distribution U which might be due to random variation.

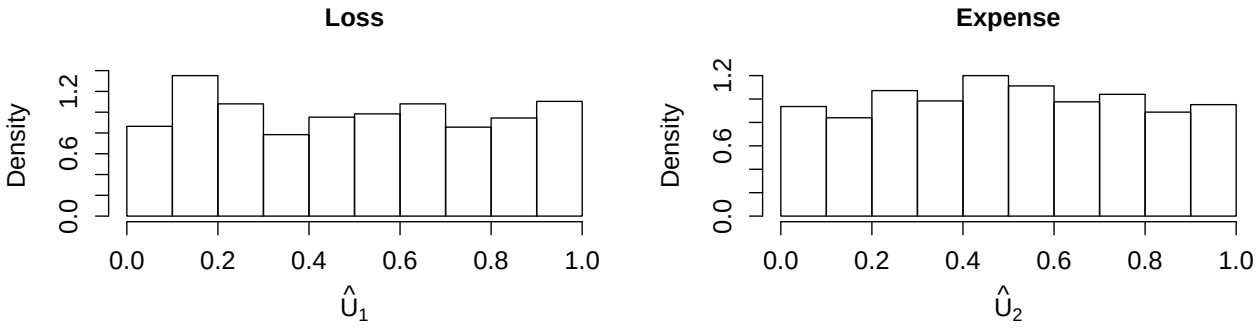

Figure 4: Histogram of the the pseudo-observations

Next, we create a bivariate scatterplot of the uniform-transformed data in Figure 5. Notice, the positive dependence stucture between `u1` and `u2`. Also the observations in the scatterpot indicate there exist upper tail dependence. An adequate model should contain a positive dependence structure and upper tail dependence. From our experience we know the Gumbel copula $C^{\mathrm{Cu}}(u_1, u_2)$ is acceptable as it contains a positive dependence stucture and upper tail dependence. The Clayton copula $C^{\mathrm{Cl}}(u_1, u_2)$ properties makes it an adequate model for lower tail dependence.

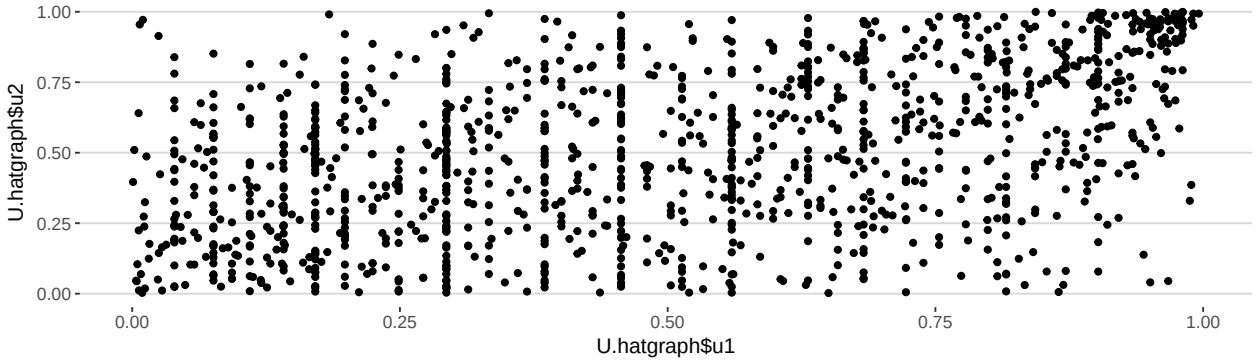

Figure 5: Bivariate scatterplot of the uniform-transformed data

Non-parametric Approach

First, we fit a Clayton and Gumbel copula using the non-parametric approach by Genest and Rivest (1993). The procedure of Genest and River (1993) provides us a strategy for selecting the parametric family of Archimedean copulas that provides the best possible fit to a given set of data. Therefore we use their estimation procedure to find the one-dimensional empirical distribution function. Figure 7 shows the empirical distribution function using the Gumbel $C^{\mathrm{Cu}}(u_1, u_2)$ and Clayton copula $C^{\mathrm{Cl}}(u_1, u_2)$. Figure 6 clearly illustrates that the non-parametric Gumbel has a better fit to the data as the non-parametric Gumbel. This result could be expected as we earlier mentioned the Gumbel will give a better fit as it incorporates the observed upper tail dependence.

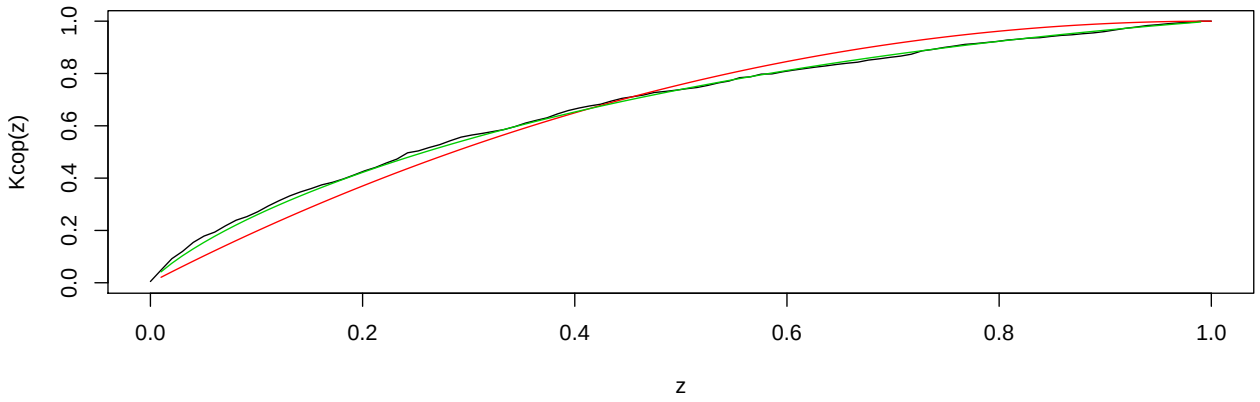

Figure 6: Archimidean Copula Identification: Non-parametric estimation (black), Non-parametric Clayton (red), Non-parametric Gumbel (green)

For the non-parametric approach we construct the dataframe `U.np`. In `U.np` we store the empirical cumulative distribution function values of the `loss` and `expense` variable using the `rank` function. First, we fit Clayton copula $C^{\text{Cl}}(u_1, u_2)$. To fit the copula to `U.np` data we will be using maximum likelihood. We use the function `fitCopula` and set the method argument equal to `ml`. Next, we fit Gumbel copula $C^{\text{Cu}}(u_1, u_2)$ using the non-parametric approach. Again we use the same data and arguments as we've done with the Clatyon copula.

Next, we plot the empircal copula `EmpCop` together with the contours of the Clayton C^{Cl} and Gumbel copula C^{Cu} we fitted using the the non-parametric approach. Figure 7 illustrates that the contours of the Gumbel copula $C^{\text{Cu}}(u_1, u_2)$ are the closest to those of the empirical copula. This can be explained by the fact that the Gumbel copula $C^{\text{Cu}}(u_1, u_2)$ is adequate for modelling upper tail dependence while the Calyton copula incorporates lower tail dependence. For this reason the Gumbel copula $C^{\text{Cu}}(u_1, u_2)$ is preferred over the Clayton copula $C^{\text{Cl}}(u_1, u_2)$ for this specific dataset.

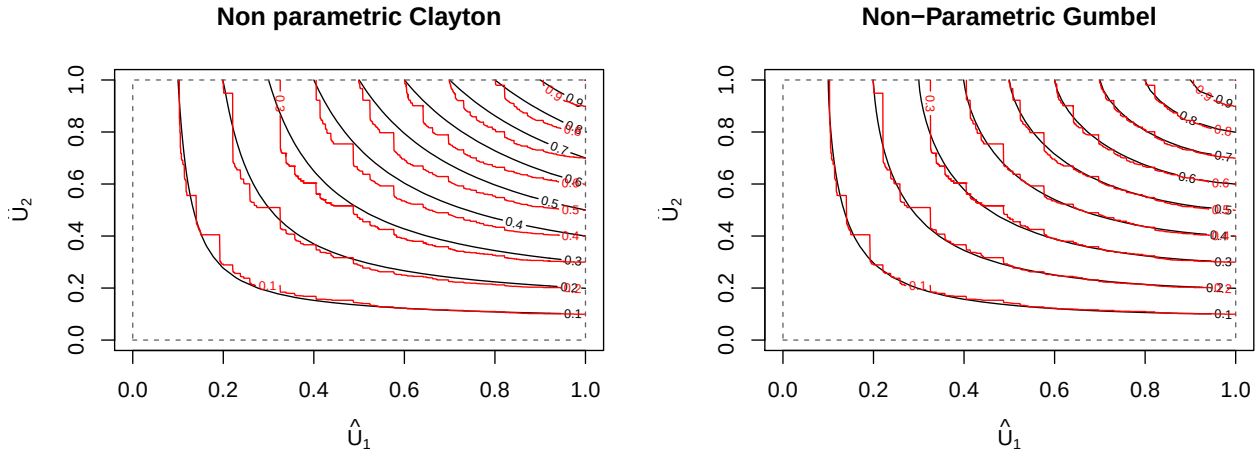

Figure 7: Empricial Copula contours (black) and Non-parametric Clayton and Gumbel (red)

Parametric Approach

Second, we fit a Gumbel copula $C^{\text{Cu}}(u_1, u_2)$ using the parametric approach. We use the maximum likelihood method. Notice, censoring is already taken into account when we fitted the Pareto model to the data.

Finally, we compare the parametric and non-parametric Gumbel $C^{\mathrm{Cu}}(u_1, u_2)$. From Figure 8 it is difficult to compare the goodness of fit. To compare all models we calculate the Akaike Information Criterium (AIC). The copula with the lowest AIC is the non-parametric Gumbel copula $C^{\mathrm{Cu}}(u_1, u_2)$. Therefore, we can say the non-parametric Gumbel copula $C^{\mathrm{Cu}}(u_1, u_2)$ gives the best fit according to the AIC. However, I would prefer the parametric Gumbel as it incorporates the censoring in the dataset. Also the AIC of the non-parametric is very close to the AIC of the parametric Gumbel. Notice, we did not take censoring into account when we use the non-parametric approach.

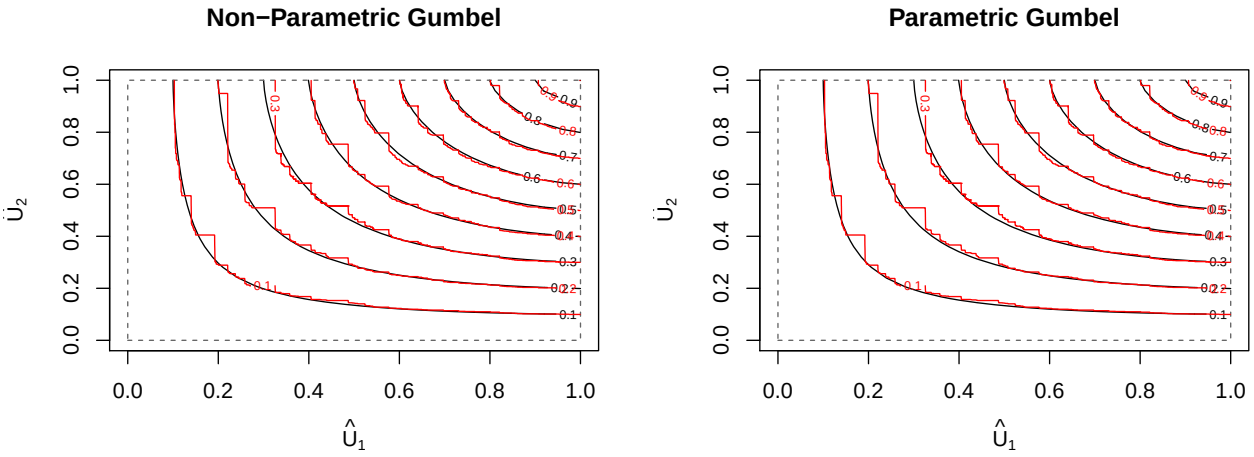

Figure 8: Empricial Copula contours (black) and non-parametric Gumbel versus Parametric Gumbel (red)

Table 4: AIC Models

	AIC
Non-Parametric Clayton	-140.7031
Non-Parametric Gumbel	-335.5985
Parametric Gumbel	-334.6358

I would defend the chosen model to my employer with a less statistical background using the contour plots. This gives an intuitive overview of the goodnes of fit for the chosen dependence model. The countour plots illustrates clearly that the model using the Gumbel copula $C^{\mathrm{Cu}}(u_1, u_2)$ has the best fit.

Finally, to generate random samples according of the chosen bivariate model for losses and expenses I would use the following procedure. I would use the function `rCopula` which can be used to generate random obsevrations given the dependence structure from our model. In this function I incorporate my chosen model and the parametric Gumbel copula $C^{\mathrm{Cu}}(u_1, u_2)$. Next, I would take the inverse of the Pareto cumulative distribution F^{-1} function to end up with bivariate observations for the `loss` and `expense` variable (see R code for more details).

References

Genest C. and Rivest L.-P. Statistical Inference Procedures for Bivariate Archimedean Copulas (1993). Journal of the American Statistical Association, 88: 1034-1043