Dependence modelling for heavy-tailed multi-peril insurance losses

by

Tianxing Yan

B.Sc., Simon Fraser University, 2022

Project Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
Department of Statistics and Actuarial Science
Faculty of Science

SIMON FRASER UNIVERSITY
Summer 2024

Declaration of Committee

Name:	**Tianxing Yan**
Degree:	**Master of Science**
Thesis title:	**Dependence modelling for heavy-tailed multi-peril insurance losses**
Committee:	**Chair:** Cherie Ng Lecturer, Statistics and Actuarial Science
	Himchan Jeong Co-supervisor Assistant Professor, Statistics and Actuarial Science
	Yi Lu Co-supervisor Professor, Statistics and Actuarial Science
	Cary (Chi-Liang) Tsai Committee Member Professor, Statistics and Actuarial Science
	Hyunwoong Chang Examiner Assistant Professor, Mathematical Sciences University of Texas at Dallas

Abstract

Assessing and managing risks is essential for insurance companies. We recognize the heavy-tailed behaviour and dependency among different coverages in insurance claim datasets. To capture claim dependency, we proposed a hierarchical model and several Copula-Based models. Composite models are applied to address the heavy-tailed behaviour of individual losses. To evaluate the performance of the proposed model from the insurance aspect, we approximate the risk measures using the Monte Carlo methodology. Finally, we demonstrate that the model considering dependency enhances model goodness-of-fit while providing more accurate risk measures of the aggregate losses for all types of coverage in total.

Keywords: Composite Model; Copula; Aggregate Loss

Dedication

My mother

Acknowledgements

I want to express my greatest gratitude to my supervisor, Professor Himchan Jeong, and my co-supervisor, Professor Yi Lu, for their sincere support and wise guidance. I have enriched and solidified my theoretical knowledge. More importantly, their long-term vision continuously supports me in reaching my academic goals.

I want to thank Cherie Ng for being willing to chair my defence. Also, I want to thank Professors Cary Tsai and Hyunwoong Chang for being committee members and examiners, for considering my work, and for providing meaningful suggestions.

I appreciate Professors Gary Parker, Jean-François Bégin, Barbara Sanders, and Cary Tsai for offering me opportunities and encouraging me to pursue an academic path. I would also like to thank all the Statistics and Actuarial Science Department faculty and staff members for their help throughout my bachelor's and master's degrees.

I also want to thank all my friends: Jinyi, Angelica, Leo, Charlie, Ken, Yuqi, Yirong, Yingfei, Rina, Vatsal, Guanzhou, Angela, Lulu, Jacob, Shifan, Yichen, Payman, Hashan, Xuefei, Boya, Kai, Dongha, Grace, Teng-Wei, Zhi Yuh, Yaning, Christy, Twist, Ian.

Finally, I want to thank the NSERC Discovery Grant and CANSSI Graduate Exchange Scholarship that supported me financially.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Insurance provides financial compensation to individuals or companies after a particular event occurs. In life insurance, the company compensates for the insured's death; in home insurance, the insurer covers the damage to the residential property. There is also a type of insurance that protects insurance companies from losses called reinsurance.

Assessing and managing potential losses are essential for insurance companies. Without a correct understanding of such risks, the insurance design might not be reasonable and may cause unexpected losses. For example, an earthquake in an area may cause substantial financial loss to the insurance companies.

Understanding the risks and developing an effective risk-managing strategy can benefit companies and the economy in several ways. First, a correct understanding of risks helps the company evaluate on a fairness basis. For instance, in life insurance, it is common for smokers to be charged higher premiums compared to nonsmokers. Next, effective risk management provides stable protection to society. Furthermore, appropriate risk management prevents the company from bankruptcy, which can cover individual losses of the insured. The previous two are the advantages of economics. Moreover, a company with a robust risk management strategy can price the insurance product competitively.

From a statistical point of view, applying random variables to transform real-life scenarios into mathematical expressions to study uncertainty is well-known. Discrete random variables are employed for countable scenarios, while continuous random variables are suitable for uncountable situations. Furthermore, selecting appropriate random variables and creating models should account for more detailed risk behaviours. Eling (2012) used several skewed distributions from the skewed-elliptical distribution family to simulate that the insurance claims are not symmetric. Hong and Martin (2018) proposed a non-parametric type of model, the Dirichlet process mixture model, which applied historical data and avoided complicated model selection processes.

1.1 Heavy-Tailed Distribution

One of the main behaviours of insurance losses is the high possibility of extreme events. During risk mitigation processes, if companies pool the risk based on the expected total claim amount, such extreme events can result in excessively large claims, dampening the solvency of the insurance portfolio. In this regard, there have been many approaches that could handle heavy-tailed behaviours of insurance claims.

Starting from the exponential distribution family, there are some heavy-tailed distributions, such as Pareto, Inverse-Gamma, and Weibull, under certain parameter settings. They often appear in actuarial literature to model the individual loss random variables.

Furthermore, two famous theorems contribute to modelling extreme values: the Fisher-Tippett-Gnedenko theorem and the Pickands-Balkema-De Haan theorem. The Fisher-Tippett-Gnedenko theorem, contributed by Fisher and Tippett (1928) and Gnedenko (1943), suggests dividing a sample into multiple subsamples and utilizes the distribution of the most significant values of subsamples is one of the Fréchet, Gumbel, or Weibull distributions. The Pickands-Balkema-De Haan theorem shows that the excess values of a sample can be studied using the generalized Pareto distribution. There has been some discussion around the two approaches. One of the most recent, Bücher and Zhou (2021), discussed the block maxima and peak-over-threshold (POT) approach in several scenarios. There are also some applications in multiple research areas. Resnick (1997) applied the POT approach to analyzing the Danish fire losses, the same dataset we used in this project; Hou and Liu (2023) investigated the mooring line for fish cages.

However, block maxima and generalized Pareto distributions only describe extreme behaviours. With a primary focus on extreme losses, the model cannot reflect other loss situations accurately. To address that, mixture and splicing models were proposed to combine multiple distributions to better describe the risk on the whole distribution's support. Miljkovic and Grün (2016) discussed several mixture models for multimodality insurance losses. Unlike the mixture models, which combine multiple distributions overlapping, the splicing models divide the whole support into multiple regions and combine distributions without overlaps. Fung et al. (2024) provided a comprehensive introduction about the mixture and splicing models and proposed soft splicing models to connect both. In this project, we apply composite models, a special type of splicing model. In particular, they combine the distribution with consideration of the continuity and differentiability at splicing points. There are quite a few articles that applied composite distributions: Scollnik and Sun (2012), Cooray and Ananda (2005), and Pigeon and Denuit (2011).

1.2 Dependency Modelling for Insurance Claims

In addition to the behaviour of a single loss, insurance companies are also interested in analyzing aggregate losses for risk diversification and company operation purposes. However, risks might be related and affect each other. Some factors affect insurance claims (or coverages) simultaneously, or different insurance lines will affect each other's likelihood of claims occurring. For instance, natural disasters in a region damage all properties and cause multiple insurance claims. Hence, hierarchical and Copula-Based modelling frameworks were regularly discussed to describe such relationships.

Hierarchical modelling bridges the relationship between different events by sharing the belief that the risks from a common environment are not independently distributed, and such modelling schemes have been widely applied in several fields. Pechon et al. (2021) introduced a correlated random effects model for multivariate credibility. It aims to study the hidden risk factors in individual policyholders that jointly affect home and motor claims. Likewise, Fung et al. (2023) applied compound claim frequency models, which simulated how storms influence the claim counts.

However, constructing such models requires a deep understanding of the data and is entirely arbitrary. Without intuitions of causation among the events, creating a model like that is challenging. Alternatively, a Copula-Based method avoids such settings by flexibly connecting random variables using a dependent structure. More specifically, we can create a joint distribution using any marginal random variables and a dependent structure. Such a dependent structure is called a copula and is defined by Sklar (1959). In his work, he proposed a theorem to decompose an existing joint distribution into marginal distributions and a copula function. Further, the theorem also allows for constructing a joint distribution from marginal distributions and a copula function. Continuing Sklar's work, several parametric copula functions have been proposed: Clayton (1978) proposed Clayton copula; Joe (1993) introduced Joe copula, etc.

The Copula-Based methodologies became more mature, extending the concept in several directions. Lee (2002) and Cameron et al. (2004) combined the copula with regression. Other than continuous margins, some research applied the copula with discrete margins. Nikoloulopoulos (2013) did inferences on the asymptotic for the multivariate normal copula with some discrete regressions. Recently, Oh et al. (2021) and Jeong et al. (2023) applied copula structure to connecting insurance claim components.

Overall, there is no winner regarding which method is the best. Hierarchical modelling is more intuitive and interpretable than Copula-Based models. However, in some scenarios, Copula-Based schemes provide more flexibility.

1.3 Motivation & Contributions

The Danish multi-peril fire loss dataset contains 2167 fire claims reported to reinsurance, which involves three loss sources: building, contents, and profits.

Over the past few decades, this dataset has been studied from different aspects.

1. McNeil (1997) applied the generalized Pareto distribution. He estimated the parameters and made inferences related to the goodness of fit.

2. Resnick (1997) studied the tail behaviour and tested the time-independent assumption for the aggregate fire loss.

3. Cabras and Castellanos (2011) implemented an additive mixture model to study the dataset. One component of the additive mixture model is a generalized Pareto distribution to capture the heavy-tailed behaviour.

Most research focused on the heavy-tailed behaviour of the aggregate loss caused by fire and provided a series of statistical inferences.

Inspired by existing research, we recognize the heavy-tailed aggregate loss in this dataset and how dependency will influence the studies' results. The impacts could be more substantial, especially when both come together. Therefore, we conduct a comprehensive study to consider both issues and discuss the results with the risk management aspect. By extending individual and collective risk models, we propose three different types of models in this project to describe the claim numbers: a Fully Independent model as a benchmark model, a hierarchical model constructed intuitively, and finally, some Copula-Based models. Meanwhile, several 2-component composite models study the severity component. After conducting statistical and risk analyses, we conclude that the models with dependency structure significantly enhance model goodness-of-fit and provide more accurate risk measures of the aggregate losses for all types of coverage in total.

The remainder of this project is organized as follows: Chapter 2 provides an overview of the modelling frameworks for frequency and severity components. In Chapter 3, we start by introducing the dataset. The proposed models are applied to Danish fire losses, followed by simulations to assess the risk measures for different proposed models. The last chapter, Chapter 4, concludes the studies.

Chapter 2

Methodology

The central intuition of modelling aggregate losses in a fixed time period is recognizing two uncertainties: the number of claims and the associated losses. Inspired by traditional and recent modelling frameworks, we study the aggregate loss in this project by constructing models that accommodate the dependency among different insurance coverages and the heavy-tailed behaviour of the losses.

In the following sections, we start with the traditional modelling framework, collective and individual risk models, for modelling the aggregate claims. Then, we introduce how the collective and individual risk models are applied in this project. Next, the maximum likelihood parameter estimation method is reviewed. Finally, in the spirit of the collective and individual risk models, we propose statistical models for both frequency and severity components.

2.1 Aggregate Loss Models

There are mainly two approaches to model aggregate claims in fixed periods, the individual and collective risk models. While the former is applied to study the loss by aggregating a certain number of individual losses, the latter is normally used to model the loss by aggregating claims occurring under one policy with a random number of claims. They are utilized depending on the purpose of analysis and applications. In our study, our proposed models aggregate the losses at the coverage level in the spirit of the collective risk model. Then, using the individual risk model, we sum the losses from three insurance coverages.

2.1.1 Individual Risk Model

Consider a group of n independent policies over a specific time period. Let S_i denote the loss incurred from the i^{th} policy during this time period, which can also be viewed as the claim amount paid to individual policy i with no insurance modification. We write $S_i = I_i \cdot Y_i$, where I_i is a Bernoulli random variable with probability p_i that a loss occurs, and Y_i is the associated loss amount with distribution function F_{Y_i}.

We can further express the distribution function of S_i for individual policy i based on the setting as

$$F_{S_i}(s_i) = \begin{cases} 1 - p_i & s_i = 0; \\ 1 - p_i + p_k \cdot F_{Y_i}(s_i) & s_i > 0, \end{cases}$$

which is a mixed distribution with a probability mass at 0 and continuous on values greater than 0. By recognizing S_i is a compound binomial distribution (i.e., $S_i = \sum_{k=1}^{I_i} Y_k$, with $S_i = 0$ when $I_i = 0$), the mean and the variance of S_i can be derived as:

$$\begin{aligned} \mathbb{E}[S_i] &= \mathbb{E}[\mathbb{E}[S_i|I_i]] \\ &= (1 - p_i) \cdot \mathbb{E}[S_i|I_i = 0] + p_i \cdot \mathbb{E}[S_i|I_i = 1] \\ &= p_i \cdot \mu_i, \\ \text{Var}[S_i] &= \mathbb{E}[\text{Var}[S_i|I_i]] + \text{Var}[\mathbb{E}[S_i|I_i]] \\ &= p_i \cdot \sigma_i^2 + p_i(1 - p_i)\mu_i^2, \end{aligned}$$

where μ_i and σ_i^2 are the mean and the variance of Y_i, respectively.

The characteristics of the aggregate loss of n independent policies, S, defined by

$$S = \sum_{i=1}^{n} S_i$$

can be obtained as follows:

- Mean: $\mathbb{E}[S] = \sum_{i=1}^{n} (p_i)\mu_i$;
- Variance: $\text{Var}[S] = \sum_{i=1}^{n} \left[p_i \left(\sigma_i^2 + (1 - p_i)\mu_i^2\right)\right]$.

Note that in a special case, where the distribution is the same for all the policies, i.e., for all $i = 1, 2, \ldots, n$, $p_i = p$, and $F_{Y_i} = F$ with $\mu_i = \mu$ and $\sigma_i^2 = \sigma^2$, the characteristics of S can be deduced to

- Mean: $\mathbb{E}[S] = \sum_{k=1}^{n} [p_k \mu_k] = np\mu$;
- Variance: $\text{Var}[S] = n[p\sigma^2 + p(1 - p)\mu^2]$.

2.1.2 Collective Risk Model

Unlike the individual risk model, the collective risk model aggregates the claims occurring in a specific time period. Let a random variable N denote the number of claims with mean μ_N and variance σ_N^2. The claim amount (or severity) random variable of the i^{th} claim is denoted by Y_i with common mean μ_Y and common variance σ_Y^2. By summing up all the losses, the corresponding aggregate loss random variable, S, can be expressed as

$$S = \sum_{i=1}^{N} Y_i.$$

Note here $S = 0$, if $N = 0$. We assume that loss random variables, $\{Y_i\}_{i \geq 1}$, are independent and identically distributed. In addition to this assumption, we assume that N is independent of all loss random variables, which allows for separately modelling frequency and severity. However, these two assumptions may not be reasonable in practice. Vernic et al. (2021) applied a bivariate Sarmoanov distribution to capture the dependence between the frequency and severity components.

Then, the attributes of the aggregate loss random variable can be derived as

$$\begin{aligned}
\mathbb{E}[S] &= \mathbb{E}\left[\mathbb{E}\left[\sum_{i=1}^{N} Y_i | N\right]\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{N} \mathbb{E}[Y_i]\right] \\
&= \mathbb{E}\left[N \cdot \mu_Y\right] \\
&= \mu_N \cdot \mu_Y, \\
\mathrm{Var}[S] &= \mathbb{E}[\mathrm{Var}[S|N]] + \mathrm{Var}[\mathbb{E}[S|N]] \\
&= \mathbb{E}[N \cdot \sigma_Y^2] + \mathrm{Var}[N \cdot \mu_Y] \\
&= \mu_N \cdot \sigma_Y^2 + \sigma_N^2 \cdot \mu_Y^2.
\end{aligned}$$

2.1.3 Discussion

As mentioned in the first chapter, our research interest is to study the Danish reinsurance aggregate losses occurring during a month or monthly aggregate losses. Specifically, we are interested in modelling total monthly losses under three insurance coverages (building, contents, and profits). The collective risk modelling framework is applied for each coverage to model the total losses. Then, the idea of individual risk framework is utilized, the add-up of the losses from three coverages is the aggregate loss we are interested in. For month t, by denoting N_{1t}, N_{2t}, and N_{3t} as the number of claims corresponding to three insurance coverages occurred in the t^{th} month, respectively, and Y_{itk} as the k^{th} claim amount under the i^{th} coverage incurred in month t, the aggregate loss of the t^{th} month, S_t, can then be expressed as

$$S_t = \sum_{i=1}^{3} \sum_{k=1}^{N_{it}} Y_{itk}. \tag{2.1}$$

In the following sections of this chapter, we introduce a Fully Independent model, a Binomial Thinning model, and three Copula-Based models. The Fully Independent model assumes no dependency among N_{1t}, N_{2t}, and N_{3t}, $t = 1, 2, \ldots$. However, the Binomial Thinning uses the number of claims reported (each claim reported could have losses in any of the three coverages) to bridge the relationships among the three claim numbers. The Copula-Based models connect three claim numbers directly using copula functions. Although the aggregate loss in the t^{th} month is the same as presented in (2.1), N_{1t}, N_{2t}, and N_{3t} are correlated under the Binomial Thinning and Copula-Based models.

2.2 Maximum Likelihood Estimations

The maximum likelihood estimation is a method that uses observations to estimate the parameters for a proposed distribution (model). This method is widely used because it is flexible and fully uses the observed data. Compared to the method of moments, which uses raw moments to estimate the model parameters, the maximum likelihood method considers every observation, which can be either a single outcome or an event.

Let $x_1, \ldots, x_T$ be T observations. Given a distribution function with a vector of parameters, Θ, the likelihood function can be written as

$$\mathcal{L}(\Theta | x_1, \ldots, x_T) = \prod_{t=1}^{T} f_X(x_t), \tag{2.2}$$

where f_X is the probability density function or probability mass function of X_is, depending on whether the distribution is continuous or discrete. However, Equation (2.2) is the likelihood function for single-value observations. The likelihood function for other types of observations, such as truncated observations, can also be constructed.

The maximum likelihood estimation provides a way to estimate the parameters of the underlying distribution, the so-called maximum likelihood estimate. The maximum likelihood estimation suggests a set of parameter values assumed model result in the observed data. In practice, one can maximize the logarithm of likelihood functions for computational and mathematical convenience. From Equation (2.2), the log-likelihood function, denoted as l, is given by

$$\begin{aligned} l(\Theta | x_1, \ldots, x_T) &= \log\left[\mathcal{L}(\Theta | x_1, \ldots, x_T)\right] \\ &= \sum_{t=1}^{T} \log\left[f_X(x_t)\right]. \end{aligned}$$

To obtain the maximum likelihood estimation of the parameters, we first take the partial derivatives of the log-likelihood function with respect to each of the parameters, and then the maximum likelihood estimators can be obtained by solving the following system of equations:

$$\begin{cases} \frac{\partial}{\partial \theta_1} l(\Theta | x_1, \ldots, x_T) = 0 \\ \qquad \vdots \\ \frac{\partial}{\partial \theta_m} l(\Theta | x_1, \ldots, x_T) = 0 \end{cases},$$

where m is the total number of parameters. In general, obtaining the closed-form solutions to this system of equations is difficult. One can apply the Newton-Raphson Method to solve such problems iteratively or use other numerical methods.

Besides determining the parameter estimates, the likelihood function values at maxima can be used to obtain a model selection criterion. Provided that the estimated parameter of a statistical model is $\hat{\Theta}$, Akaike information and Bayesian information criteria (AIC and BIC) are widely used and defined as follows:

$$\begin{aligned} \text{AIC} &= 2m - 2l(\hat{\Theta} | x_1, \ldots, x_T), \\ \text{BIC} &= m \ln(T) - 2l(\hat{\Theta} | x_1, \ldots, x_T), \end{aligned}$$

where m represents the number of parameters (the number of elements of $\hat{\Theta}$). AIC is a model evaluation criterion that considers the model's complexity into account. Given the same log-likelihood function values under two different models, AIC suggests using a model with fewer parameters. BIC considers both the number of parameters and the number of observations. Both AIC and BIC select the most preferred model by choosing the smallest value.

2.3 Frequency Modelling Frameworks

In this section, we propose three types of frequency models to study the number of claims: Fully Independent model, Binomial Thinning model, and Copula-Based models. The independent model is used as a benchmark model for comparison. The Binomial Thinning model is a typical model that captures the dependency among the claim numbers in different lines. However, it only allows a fixed dependency structure between any two margins. Copula-Based models construct the joint distribution of claim counts from different lines of business with great flexibility. These three types of models are to be introduced respectively in the following three subsections.

2.3.1 Benchmark Model: Independent Frequency Model

From Equation (2.1), by directly applying collective and individual risk models, we get a Fully Independent compound model, which is treated as a benchmark model throughout the remaining analyses. In addition to the independence assumption between the frequency and severity components, it also assumes that the number of claims under different coverages are independently distributed.

Let N_{1t}, N_{2t} and N_{3t} represent the number of claims under the building, contents, and profits coverages, respectively, in the t^{th} month. All claim numbers observed from the three coverages are overdispersed. Because of that, we assume that $N_{it} \sim \mathcal{NB}(r_i, \lambda_i)$ for the i^{th} insurance coverage. The probability mass function of claim numbers N_{it} is

$$\Pr(N_{it} = n_{it}) = \frac{\Gamma(r_i + n_{it})}{\Gamma(r_i)\Gamma(n_{it}+1)} \left(\frac{\lambda_i}{r_i + \lambda_i}\right)^{n_{it}} \left(\frac{r_i}{r_i + \lambda_i}\right)^{r_i}, \qquad n_{it} = 0, 1, 2, \ldots, \tag{2.3}$$

with mean and variance being λ_i and $\lambda_i(1 + \lambda_i/r_i)$.

The likelihood function with T (months) observations can be written as

$$\begin{aligned} \mathcal{L}(\Theta|\mathcal{D}) &= \prod_{t=1}^{T} \Pr(N_{1t} = n_{1t}, N_{2t} = n_{2t}, N_{3t} = n_{3t}; \Theta) \\ &\overset{\text{ind.}}{=} \prod_{t=1}^{T} \Pr(N_{1t} = n_{1t}; \Theta) \cdot \Pr(N_{2t} = n_{2t}; \Theta) \cdot \Pr(N_{3t} = n_{3t}; \Theta) \\ &= \prod_{t=1}^{T} \prod_{i=1}^{3} \left[\frac{\Gamma(r_i + n_{it})}{\Gamma(n_{it}+1)\Gamma(r_i)} \frac{(\lambda_i/r_i)^{n_{it}}}{(1 + \lambda_i/r_i)^{r_i + n_{it}}} \right], \end{aligned} \tag{2.4}$$

where Θ is a vector of six parameters, and $\mathcal{D}$ is an observation set which contains the observed claim numbers of all three coverages for T months.

For parameter estimations, we use the maximum likelihood approach and use maximum likelihood estimators for further investigation. Because of the independent assumption, we do not need to start with the full likelihood. Instead of (2.4), we can estimate the parameters of each marginal distribution separately. In this case, the likelihood function for claim numbers of i^{th} coverage is given by

$$\begin{aligned} \mathcal{L}_i(r_i, \lambda_i | n_{i1}, \ldots, n_{iT}) &= \prod_{t=1}^{T} \Pr(N_{it} = n_{it}; r_i, \lambda_i) \\ &= \prod_{t=1}^{T} \frac{\Gamma(r_i + n_{it})}{\Gamma(r_i)\Gamma(n_{it}+1)} \left(\frac{\lambda_i}{r_i + \lambda_i}\right)^{n_{it}} \left(\frac{r_i}{r_i + \lambda_i}\right)^{r_i}. \end{aligned}$$

Then, its log-likelihood can be expressed as

$$\begin{aligned} l_i(r_i, \lambda_i | n_{i1}, \ldots, n_{iT}) &= \sum_{t=1}^{T} \log[\Pr(N_i = n_i; r_i, \lambda_i)] \\ &= \sum_{t=1}^{T} \log[\Gamma(r_i + n_{it})] - T\log[\Gamma(r_i)] - \sum_{t=1}^{T} \log[\Gamma(n_{it}+1)] \\ &\quad + \sum_{t=1}^{T} n_{it} \log\left(\frac{\lambda_i}{r_i + \lambda_i}\right) + T r_i \log\left(\frac{r_i}{r_i + \lambda_i}\right). \end{aligned} \tag{2.5}$$

Since the maximum likelihood estimators of the negative binomial distribution do not have closed-form expressions, we apply a numerical method to approximate the parameter values that maximize the log-likelihood given by (2.5) for $i = 1, 2, 3$.

2.3.2 Binomial Thinning Model

Hierarchical models inspire us to propose a joint frequency distribution at a multilevel since the data contains the number of claims reported each month, and each claim reported could have losses in any of the three insurance coverages. Given the number of claims reported in month t, denoted as M_t, it is natural to assume that the claim numbers for different insurance coverages are all binomially distributed with size parameter M_t, and the three claim counting random variables are conditionally independent.

We propose a Binomial Thinning model, in which we assume a negative binomial model for the reported claim numbers with parameters r and λ. Furthermore, we assume that for $i = 1, 2, 3$, $(N_{it}|M_t) \sim \mathcal{BN}(M_t, \lambda_i/\lambda)$, where λ_i/λ is the probability of a reported claim having a loss in i^{th} coverage, for the three sources of claim numbers. The joint probability mass function can be expressed as

$$\begin{aligned} &\Pr(M_t = m_t, N_{1t} = n_{1t}, N_{2t} = n_{2t}, N_{3t} = n_{3t}) \\ &= \Pr(M_t = m_t) \cdot \Pr(N_{1t} = n_{1t} | M_t = m_t) \cdot \Pr(N_{2t} = n_{2t} | M_t = m_t) \cdot \Pr(N_{3t} = n_{3t} | M_t = m_t) \\ &= \left[\frac{\Gamma(r + m_t)}{\Gamma(m_t + 1)\Gamma(r)} \frac{(\lambda/r)^{m_t}}{(1 + \lambda/r)^{r + m_t}}\right] \cdot \prod_{i=1}^{3} \left[\binom{m_t}{n_{it}} \frac{\lambda_i^{n_{it}} (\lambda - \lambda_i)^{m_t - n_{it}}}{\lambda^{m_t}}\right]. \end{aligned} \tag{2.6}$$

Since the three claim numbers are conditionally independent of each other, given the number of reported claims, we can obtain that the marginal distribution of N_{it} is a negative binomial with size parameter r and mean λ_i. The detailed derivation is shown below:

$$\begin{aligned}
&\Pr(N_{it} = n_{it}) \\
&= \sum_{m_t=n_{it}}^{\infty} [\Pr(N_{it} = n_{it}|M_t = m_t) \cdot \Pr(M_t = m_t)] \\
&= \sum_{m_t=n_{it}}^{\infty} \left[\frac{\Gamma(r+m_t)}{\Gamma(m_t+1)\Gamma(r)} \frac{(\lambda/r)^{m_t}}{(1+\lambda/r)^{r+m_t}}\right] \cdot \left[\binom{m_t}{n_{it}} \frac{\lambda_i^{n_{it}}(\lambda-\lambda_i)^{\lambda-n_{it}}}{\lambda^{m_t}}\right] \\
&= \frac{\Gamma(r+n_{it})}{\Gamma(n_{it}+1)\Gamma(r)} \sum_{m_t=n_{it}}^{\infty} \binom{r+m_t-1}{m_t-n_{it}} \left(\frac{\lambda-\lambda_i}{\lambda}\right)^{m_t-n_{it}} \left(\frac{\lambda_i}{\lambda}\right)^{n_{it}} \left(\frac{\lambda}{\lambda+r}\right)^{m_t} \left(\frac{r}{\lambda+r}\right)^{r} \\
&= \frac{\Gamma(r+n_{it})}{\Gamma(n_{it}+1)\Gamma(r)} \left(\frac{\lambda_i}{\lambda_i+r}\right)^{n_{it}} \left(\frac{r}{\lambda_i+r}\right)^{r} \sum_{m_t=n_{it}}^{\infty} \binom{r+m_t-1}{m_t-n_{it}} \frac{(\lambda_i+r)^{n_{it}+r}(\lambda-\lambda_i)^{m_t-n_{it}}}{(\lambda+r)^{m_t+r}} \\
&= \frac{\Gamma(r+n_{it})}{\Gamma(n_{it}+1)\Gamma(r)} \left(\frac{\lambda_i}{\lambda_i+r}\right)^{n_{it}} \left(\frac{r}{\lambda_i+r}\right)^{r}, \qquad n_{it} = 0, 1, 2, \ldots.
\end{aligned}$$

Note that the claim numbers under each coverage cannot exceed the reported claims. As a result, the summation for m_t in the above derivation is summing from n_{it} instead of 0. We can write the last step directly from the previous step because we recognize that the summation is summing probability mass functions of a negative binomial from 0 to infinity. An alternative way to show that the marginal distribution of claim numbers is negative binomial is to use generating functions, either moment-generating or characteristic functions.

The likelihood function for T months data is the product of T joint probability mass functions, given by (2.6). It can be written as

$$\begin{aligned}
\mathcal{L}(\Theta|\mathcal{D}) &= \prod_{t=1}^{T} \Pr(M_t = m_t, N_{1t} = n_{1t}, N_{2t} = n_{2t}, N_{3t} = n_{3t}), \\
&= \prod_{t=1}^{T} \left\{ \left[\frac{\Gamma(r+m_t)}{\Gamma(m_t+1)\Gamma(r)} \frac{(\lambda/r)^{m_t}}{(1+\lambda/r)^{r+m_t}}\right] \cdot \prod_{i=1}^{3} \left[\binom{m_t}{n_{it}} \frac{\lambda_i^{n_{it}}(\lambda-\lambda_i)^{m_t-n_{it}}}{\lambda^{m_t}}\right] \right\}. \qquad (2.7)
\end{aligned}$$

Compared with the independent model, this Binomial Thinning considers the dependency among three lines of business. However, an obvious drawback is the unchangeable dependent structure. The dependent relationship is tied to the marginal distributions. For the Binomial Thinning model, marginals are consistently correlated. If the observation shows a more correlated dependency in extreme cases, such hierarchical models cannot flexibly adjust the relationship among the marginal random variables.

2.3.3 Copula-Based Frequency Model

To overcome the drawback of the Binomial Thinning model, we consider the copula approach. Copula is a joint distribution with standard uniform marginal distributions. By treating the cumulative distribution function as a standard uniform random variable, one can combine any existing continuous marginal distributions with a copula function and get a new joint distribution. This process can be reversed. The idea was first proposed by Sklar (1959) and applied in different research areas.

This section reviews the basic concept of copula, its definition, and its associated properties. Moreover, some implicit and explicit copulas are introduced as examples that will be applied in the application chapter. We also discuss applying copula with discrete random variables and the likelihood function.

Sklar (1959) defined copula functions. The main theorem is stated below.

Theorem 1. *Let $X_1, \ldots, X_n$ be n continuous random variables. Their corresponding distributions are denoted by $F_1, \ldots, F_n$. The corresponding copula function is*

$$C(F_1(x_1), \ldots, F_n(x_n)) = F(x_1, \ldots, x_n), \tag{2.8}$$

where F is the joint distribution of X_is.

By using Theorem 1, we can split a joint distribution into two parts. One part is the marginal distributions, and the other is the copula function, which we treat as a built-in dependent structure of the joint distribution. In addition, we can build a joint distribution by combining the marginal distributions with a copula function to capture the dependence between the marginals.

We can get its density since we can treat a copula function as a joint cumulative distribution of n standard uniform random variables.

Corollary 1.1. *Given that $C(F_1(x_1), \ldots, F_n(x_n))$ defined by (2.8) is a joint distribution function, the corresponding joint density for n uniform margins and the joint density for $x_1, \ldots, x_n$ are*

$$\begin{aligned} c(F_1(x_1), \ldots, F_n(x_n)) &= \frac{\partial^n}{\partial F_1(x_1) \cdots \partial F_n(x_n)} \left[C(F_1(x_1), \ldots, F_n(x_n)) \right], \\ f(x_1, \ldots, x_n) &= \frac{\partial^n}{\partial x_1 \cdots \partial x_n} \left[C(F_1(x_1), \ldots, F_n(x_n)) \right] \\ &= f_1(x_1) \cdots f_n(x_n) \cdot c((F_1(x_1), \ldots, F_n(x_n))), \end{aligned} \tag{2.9}$$

where c is the joint probability density function of $F_1(X_1), \ldots, F_n(X_n)$, f is the joint probability density function of X_is corresponding to F, and $\forall i = 1, \ldots, n$, f_i are marginal density functions.

The following can be proved since copula functions are joint distribution functions.

Corollary 1.2. *Suppose that $C(F_1(x_1), \ldots, F_n(x_n))$ is a copula function, $0 \leq u_1 \leq u_2 \leq 1$. For all $j \in \{1, \ldots, n\}$, the following properties hold:*

- $C(0,\ldots,0)=0$;
- $C(1,\ldots,1)=1$;
- $C(F_1(x_1),\ldots,F_j(u_1),\ldots,F_n(x_n))\leq C(F_1(x_1),\ldots,F_j(u_2),\ldots,F_n(x_n))$;
- $C(F_1(x_1),\ldots,0,\ldots,F_n(x_n))=0$;
- $C(1,\ldots,F_j(x_1),\ldots,1)=F_j(x_1)$.

Figure 2.1 are examples of two-dimensional copulas. The left and right panels are the 2-dimensional lower and upper Fréchet–Hoeffding copula bounds. They indicate the perfectly negative and positive relationships, respectively, between two random variables. All copula functions are bounded by these two surfaces. The middle panel is a Gaussian copula (see below for its definition).

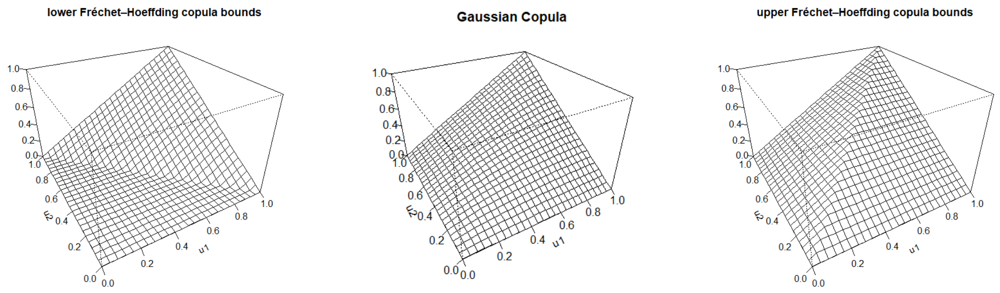


Figure 2.1: Three Copulas

There are two main types of copula functions: implicit and explicit. Implicit copulas are extracted from the existing multivariate distributions. However, they are not required to have explicit expressions. Unlike implicit copulas, explicit copulas have explicit formulas, for which they are constructed by using generating functions.

By Theorem 1, a way to generate a dependent relationship is to extract it from the existing joint distributions. Such types of copula are called implicit copulas. A Gaussian copula is extracted from a multivariate normal distribution. Let $U_1,\ldots,U_n\sim\mathcal{U}(0,1)$ with correlation matrix P.

$$C^n_{Gauss}(u_1,\ldots,u_n)=\Phi_P(\Phi^{-1}(u_1),\ldots,\Phi^{-1}(u_n)),\qquad(2.10)$$

where Φ^{-1} is the inverse of the cumulative distribution function of a standard normal random variable, Φ_P is the cumulative distribution function of multivariate standard normal with correlation or covariance matrix P. The parameters associated are elements in the correlation matrix, which has $n(n-1)/2$ parameters describing the pairwise correlations.

Given a pairwise correlation is equal to 1, the pair are perfectly positively correlated to each other. The pair are perfectly negatively correlated if the pairwise correlation is -1.

In addition to the copulas extracted from existing distributions, a widely used class of explicit copulas is Archimedean copulas. It provides a variety of dependent structures and can be easily applied in practice.

Definition 1. *An Archimedean generator ψ is a strictly decreasing, concave, and continuous function. It maps numbers in $(0, \infty)$ to $[0, 1]$ with $\psi(0) = 1$ and $\psi(\infty) = 0$. The corresponding Archimedean copula function is defined by*

$$C(u_1, \ldots, u_n) = \psi(\psi^{-1}(u_1) + \cdots + \psi^{-1}(u_n)).$$

We now show two Archimedean copula structures, Gumbel and Joe copulas. The Gumbel copula, introduced by Gumbel (1960), is constructed by the generator, $\psi_G(t) = \exp(-t^{1/\theta})$, for $t \geq 0$ and $\theta \geq 1$. By the definition of the Archimedean copulas, the Gumbel copula is defined as

$$\begin{aligned} C_G^n(u_1, \ldots, u_n) &= \psi_G\left(\psi_G^{-1}(u_1) + \cdots + \psi_G^{-1}(u_n)\right) \\ &= \exp\left[-(-(\log u_1)^\theta - \cdots - (\log u_n)^\theta)^{1/\theta}\right], \qquad u \in [0, 1], \end{aligned} \tag{2.11}$$

where $\theta \geq 1$. The parameter can be interpreted as a level of dependency. When $\theta = 1$, there is no dependency among the margins; when θ increases, the tails of margins become more and more positively correlated. The Joe copula is proposed by Joe (1993) with the generator $\psi_J(t) = 1 - [1 - \exp(-t)]^{1/\theta}$, for $t \geq 0$ and $\theta \geq 1$. The Joe copula can be written as

$$\begin{aligned} C_J^n(u_1, \ldots, u_n) &= \psi_J\left(\psi_J^{-1}(u_1) + \cdots + \psi_J^{-1}(u_n)\right) \\ &= 1 - \left[1 - \exp\left(-\sum_{j=1}^{n}\left(-\ln\left(1 - (1 - u_j)^\theta\right)\right)\right)\right]^{1/\theta}, \qquad u \in [0, 1], \end{aligned} \tag{2.12}$$

where $\theta \geq 1$. A larger θ indicates that the margins are more positively correlated. There are many other examples of Archimedean copulas, such as Clayton copula introduced by Clayton (1978).

We now consider the maximum likelihood estimation in the copula model situation. With the observed information set $\mathcal{D} = \{x_{it}, i = 1, \ldots, n, t = 1, \ldots, T\}$, and using the joint density function, (2.9), the likelihood and log-likelihood functions can be expressed as

$$\begin{aligned}
\mathcal{L}(\Theta|\mathcal{D}) &= \prod_{t=1}^{T} f(x_{1t}, \ldots, x_{nt}) \\
&= \prod_{t=1}^{T} f_1(x_{1t}) \cdots f_n(x_{nt}) \cdot c((F_1(x_{1t}), \ldots, F_n(x_{nt}))), \\
l(\Theta|\mathcal{D}) &= \sum_{t=1}^{T} \log f(x_{1t}, \ldots, x_{nt}) \\
&= \sum_{t=1}^{T} \left[\log f_1(x_{1t}) + \cdots + \log f_n(x_{nt})\right] + \sum_{t=1}^{T} \left[\log c((F_1(x_{1t}), \ldots, F_n(x_{nt}))\right], \qquad (2.13)
\end{aligned}$$

where Θ is a vector of parameters of the marginal distributions and the copula. The procedure of estimating the parameters uses the first part of Equation (2.13) as a profile likelihood for estimating the parameters of the marginal distributions. Then, by plugging in the estimated parameter values, we determine the maximum likelihood estimators of copula parameters that maximize the log-likelihood function given by Equation (2.13).

To apply copula with discrete random variables, we now introduce a way to transform the copula density into a mass function for discrete random variables. Let X_1, X_2, and X_3 be discrete random variables with supports that are non-negative integers with cumulative distribution functions F_1, F_2, and F_3, respectively.

$$\begin{aligned}
c(F_1(x_1), F_2(x_2)) &= \Pr[X_1 = x_1, X_2 = x_2] \\
&= \Pr[x_1 - 1 < X_1 \leq x_1, x_2 - 1 < X_2 \leq x_2] \\
&= \Pr(X_1 \leq x_1, X_2 \leq x_2) \\
&\quad - \Pr(X_1 \leq x_1 - 1, X_2 \leq x_2) - \Pr(X_1 \leq x_1, X_2 \leq x_2 - 1) \\
&\quad + \Pr(X_1 \leq x_1 - 1, X_2 \leq x_2 - 1) \\
&= C(F_1(x_1), F_2(x_2)) \\
&\quad - C(F_1(x_1 - 1), F_2(x_2)) - C(F_1(x_1), F_2(x_2 - 1)) \\
&\quad + C(F_1(x_1 - 1), F_2(x_2 - 1)).
\end{aligned}$$

Similarly, for a three-dimensional copula function, the corresponding copula mass function can be derived as

$$\begin{aligned} c(F_1(x_1), F_2(x_2), F_3(x_3)) = {} & C(F_1(x_1), F_2(x_2), F_3(x_3)) \\ & - \sum_{h_1=1}^{3} C(\dots, F_{h_1}(x_{h_1} - 1), \dots) \\ & + \sum_{\substack{h_1, h_2 \in \{1,2,3\}; \\ h_1 \neq h_2}}^{3} C(\dots, F_{h_1}(x_{h_1} - 1), F_{h_2}(x_{h_2} - 1), \dots) \\ & - \sum_{\substack{h_1, h_2, h_3 \in \{1,2,3\}; \\ h_1 \neq h_2 \neq h_3}}^{3} C(F_{h_1}(x_{h_1} - 1), F_{h_2}(x_{h_2} - 1), F_{h_3}(x_{h_3} - 1)). \end{aligned} \tag{2.14}$$

There are concerns regarding applying copula with discrete random variables. Instead of a unique copula function with continuous margins, the discrete one may have an infinite number of copula functions that can match the joint distribution. Assume we have two dependent Bernoulli random variables. The sample space $\Omega = \{(0,0), (0,1), (1,0), (1,1)\}$ with probabilities $0.2, 0.3, 0.1$, and 0.4, respectively. It is evident that determining a copula function with a small sample space is not robust enough. However, there are still some reasons to use a copula with discrete random variables. Instead of two possible random values for each margin, consider margins with supports being all non-negative integers such as negative binomial. With a larger sample space, the possible copula function will be more restricted and more robust when we compare different copula structures. Although there are still infinite surfaces that pass through all the points, applying different surfaces will not drastically affect the joint distribution with a specific shape of the surface.

In this project, we combine three negative binomial random variables with each of the three copula functions: Gaussian, Gumbel, and Joe copulas. Let N_{1t}, N_{2t}, N_{3t} be negatively binomially distributed with $N_{it} \sim \mathcal{NB}(r_i, \lambda_i), i = 1, 2, 3$. The log-likelihood function of a Copula-Based frequency component can be derived from Equation (2.13) by plugging in the probability mass functions of the margins, (2.3). The second part of Equation (2.13) is obtained by discretizing the copula functions, (2.10), (2.11), and (2.12) using (2.14).

2.4 Severity Modelling Framework

We have considered the models for the frequency component of the aggregate loss model in the last section. In this section, we study the models for the severity component. As mentioned in the first chapter, short-term insurance loss data is usually heavy-tailed. Composite models can capture this behaviour while maintaining a good fit for the head part of the loss distribution.

A 2-component composite model combines the head part of a light-tailed distribution with the tail part of a heavy-tailed distribution. Denote the densities for the head part and tail part as $g_1(Y)$ and $g_2(Y)$ with cumulative distribution functions $G_1(Y)$ and $G_2(Y)$, respectively. The density function of the 2-component composite model for the loss/claim amount random variable can be expresses as

$$g_{comp}(y) = \begin{cases} \frac{1}{1+\phi}\, g_1^*(y), & y < u; \\ \frac{\phi}{1+\phi}\, g_2^*(y), & y \geq u, \end{cases} \tag{2.15}$$

where ϕ is a weight parameter, u represents the threshold that separates the two components, and g_1^* and g_2^* denote the truncated distributions, given by $g_1^*(y) = g_1(y)/G_1(u)$ and $g_2^*(y) = g_2(y)/(1 - G_2(u))$. The cumulative distribution function of this composite model can be expressed as

$$\begin{aligned} G_{comp}(y) &= \begin{cases} \frac{1}{1+\phi}\, G_1^*(y), & y < u; \\ \frac{1}{1+\phi} + \frac{\phi}{1+\phi}\, G_2^*(y), & y \geq u, \end{cases} \\ &= \begin{cases} \frac{1}{1+\phi}\, \frac{G_1(y)}{G_1(u)}, & y < u; \\ \frac{1}{1+\phi} + \frac{\phi}{1+\phi}\, \frac{G_2(y)}{1-G_2(u)}, & y \geq u. \end{cases} \end{aligned}$$

where G_1^* and G_2^* are the cumulative distribution function of truncated random variables corresponding to g_1^* and g_2^*. The inverse of the cumulative distribution function can be obtained as

$$G_{comp}^{-1}(p) = \begin{cases} G_1^{-1}\left(p \cdot (1+\phi) \cdot G_1(u)\right), & p < \frac{1}{1+\phi}; \\ G_2^{-1}\left(\left(\frac{p(1+\phi)}{\phi} - \frac{1}{\phi}\right) \cdot (1 - G_2(u))\right), & p \geq \frac{1}{1+\phi}. \end{cases} \tag{2.16}$$

We only need to estimate the parameters of the head and tail distributions. In this study, the threshold and weight parameters are determined by assuming that the density function of this composite model is continuous and differentiable at the threshold. More specifically, by the continuity assumption at u for (2.15), we have $\lim_{y\to u^-} g(y) = \lim_{y\to u^+} g(y)$, which can be induced to obtain

$$\phi = \frac{\lim_{y\to u^-} g_1^*(y)}{\lim_{y\to u^+} g_2^*(x)} = \frac{g_1(u)(1 - G_2(u))}{g_2(u)G_1(u)}. \tag{2.17}$$

From the differentiability aspect for density function (2.15) at u, the following equality holds:

$$\frac{1}{1+\phi} \lim_{y\to u^-} \frac{\mathrm{d}}{\mathrm{d}y} g_1^*(y) = \frac{\phi}{1+\phi} \lim_{y\to u^+} \frac{\mathrm{d}}{\mathrm{d}y} g_2^*(y). \tag{2.18}$$

We further plug in the weight parameter given by Equation (2.17), and then Equation (2.18) can be rewritten as

$$\frac{\mathrm{d}}{\mathrm{d}u}\ln\left(\frac{g_1(u)}{g_2(u)}\right)=0. \tag{2.19}$$

The following is an example to illustrate the process of determining the threshold and weight parameters for Gamma & Inverse-Gamma composite models, given that we know the distribution parameters. Let $Y_1 \sim \mathcal{G}amma(\alpha_1,\theta_1)$ and $Y_2 \sim Inv - \mathcal{G}amma(\alpha_2,\theta_2)$. Their density functions are given by

$$\begin{aligned} g_1(y_1) &= \frac{(y_1/\theta_1)^{\alpha_1}e^{-y_1/\theta_1}}{y_1\Gamma(\alpha_1)}, \qquad y_1 \geq 0,\\ g_2(y_2) &= \frac{(\theta_2/y_2)^{\alpha_2}e^{-\theta_2/y_2}}{y_2\Gamma(\alpha_2)}, \qquad y_2 \geq 0. \end{aligned}$$

By using Equations (2.17) and (2.19), we can obtain the threshold and weight parameters as functions of the distribution parameters. In this case,

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}u}\left[\ln\frac{g_1(u)}{g_2(u)}\right]\\ =&\frac{\mathrm{d}}{\mathrm{d}u}\left[\ln\frac{\frac{(u/\theta_1)^{\alpha_1}e^{-u/\theta_1}}{u\Gamma(\alpha_1)}}{\frac{(\theta_2/u)^{\alpha_2}e^{-\theta_2/u}}{u\Gamma(\alpha_2)}}\right]\\ =&\frac{\mathrm{d}}{\mathrm{d}u}\left[\alpha_1\ln u-\frac{u}{\theta_1}+\alpha_2\ln u+\frac{\theta_2}{u}\right]\\ =&(\alpha_1+\alpha_2)\frac{1}{u}-\frac{1}{\theta_1}-\frac{\theta_2}{u^2}. \end{aligned}$$

Then, we can use the u obtained to get the weight parameter ϕ using (2.17).

Table 2.1 lists the expression of the threshold parameter u or an equation to determine u for other composite models that we consider in this project. Besides, one should recognize that calculating the weight parameter can always be done by using (2.17).

Name	Head Dist.	Tail Dist.	u
G & IG	$\frac{(x/\theta_1)^{\alpha_1}e^{-x/\theta_1}}{x\Gamma(\alpha_1)}$	$\frac{(\theta_2/x)^{\alpha_2}e^{-\theta_2/x}}{x\Gamma(\alpha_2)}$	$u=\frac{(\alpha_1+\alpha_2)+\sqrt{(\alpha_1+\alpha_2)^2-4\frac{\theta_2}{\theta_1}}}{2/\theta_1}$
G & LN	$\frac{(x/\theta_1)^{\alpha_1}e^{-x/\theta_1}}{x\Gamma(\alpha_1)}$	$\frac{\exp\left\{-\frac{(\ln x-\mu_2)^2}{2\sigma_2^2}\right\}}{x\sigma_2\sqrt{2\pi}}$	$0=\alpha_1-\frac{u}{\theta_1}+\frac{\ln x-u}{\sigma_2^2}$
G & Pa	$\frac{(x/\theta_1)^{\alpha_1}e^{-x/\theta_1}}{x\Gamma(\alpha_1)}$	$\frac{\alpha_2\theta_2^{\alpha_2}}{(x+\theta_2)^{\alpha_2+1}}$	$u=\frac{(\alpha_1+\alpha_2-\frac{\theta_2}{\theta_1})+\sqrt{(\alpha_1+\alpha_2\frac{\theta_2}{\theta_1})^2+4\frac{\theta_2}{\theta_1}(\alpha_1-1)}}{2/\theta_1}$
E & IG	$\frac{e^{-x/\theta_1}}{\theta_1}$	$\frac{(\theta_2/x)^{\alpha_2}e^{-\theta_2/x}}{x\Gamma(\alpha_2)}$	$u=\frac{(\alpha_2+1)+\sqrt{(\alpha_2+1)^2-4\frac{\theta_2}{\theta_1}}}{2/\theta_1}$
E & LN	$\frac{e^{-x/\theta_1}}{\theta_1}$	$\frac{\exp\left\{-\frac{(\ln x-\mu_2)^2}{2\sigma_2^2}\right\}}{x\sigma_2\sqrt{2\pi}}$	$0=-\frac{1}{\theta_1}+\frac{1}{u}+\frac{\ln u-\mu_2}{u\sigma_2^2}$
E & Pa	$\frac{e^{-x/\theta_1}}{\theta_1}$	$\frac{\alpha_2\theta_2^{\alpha_2}}{(x+\theta_2)^{\alpha_2+1}}$	$u=(\alpha_2+1)\theta_1-\theta_2$

Table 2.1: Composite Models

For the first column of the table, we use abbreviations to represent different distributions. G & IG indicate the Gamma & Inv-Gamma composite model where Gamma distribution represents modelling the head part of the composite distribution, and Inverse-Gamma stands to model the tail part. Besides, LN represents Log-Normal; Pa stands for Pareto; E is used for the head part and represents Exponential distribution. The Head Dist. and Tail Dist. columns show the parameterization of the distributions of head and tail parts.

Chapter 3

Application in Danish Reinsurance Dataset

In this chapter, the frequency and severity models that we propose in Chapter 2 are applied to the Danish reinsurance dataset for risk analyses of the aggregate loss under all three insurance coverages. We consider the composite model to capture the heavy-tailed behaviour in the dataset. Meanwhile, by comparing the frequency models through risk analyses, we conclude that a relationship among different insurance coverages in this dataset exists and is influential.

In the following sections, we first show our data exploration results. We then fit the data to our proposed frequency and severity models. Finally, we apply the Monte Carlo approach to approximate the risk measures under different models that we study in this project.

3.1 Data Exploration

In this project, we consider a well-known dataset, the Danish multi-peril fire losses, which is available in R library, "CASdataset", and recorded by Copenhagen Reinsurance company, which contains 2167 fire loss records in Danish Krone from 1980 to 1990. Each recorded claim includes the loss amounts of three sections: building, contents, and profits. Table 3.1 shows a few rows of the data provided by this dataset. The Building, Contents, and Profits columns show the Danish Krone losses in the millions, adjusted by inflation using 1985 as the base year. The Total column aggregates the three amounts.

Date	Building	Contents	Profits	Total
1980-01-03	1.09809663	0.58565150	0.00000000	1.683748
1980-01-04	1.75695461	0.33674960	0.00000000	2.093704
1980-01-05	1.73258126	0.00000000	0.00000000	1.732581
1980-01-07	0.00000000	1.30537600	0.47437775	1.779754
1980-01-07	1.24450952	3.36749600	0.00000000	4.612006

Table 3.1: Excerpt from the Danish Fire Dataset

As mentioned in Chapter 2, our focus is on modelling the aggregate losses during a specific time period; here, we consider the monthly aggregate losses. Before applying the modelling technique, we aggregate the claim numbers on a monthly basis to obtain the following observations: the number of claims for each month, and the associated loss amount under the i^{th} line of insurance for each claim, for $i = 1, 2, 3$, where $i = 1$ represents the damage to the building, $i = 2$ is the contents related, and $i = 3$ stands for the loss in profits.

After reorganizing the original data, we have observations for a total of 132 months. Table 3.2 summarizes the monthly number of claims. The possible claim numbers range from 0 to infinity. We also found overdispersion behaviours for all claim numbers. Based on these, we conclude to use a negative binomial random variable for claim numbers. From the summary statistics, all four claim numbers have means that are close to their medians. In addition, the number of claims under building coverage has the highest mean value among the three coverages. Losing profits are the least likely to happen. Besides, recall that a claim reported will lead to losses in any of the three insurance coverages. Therefore, the maximum of the reported claim number is the largest compared to the other three coverages' claim numbers. These behaviours on the claim numbers can also be explored from the boxplots shown in Figure 3.1.

Source	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Building	4	11	15	15.08	18	34
Contents	1	10	13	12.73	15	31
Profits	0	3	4	4.67	6.25	19
Claims Reported	7	13	16	16.42	19.25	37

Table 3.2: Summary of Claim Numbers

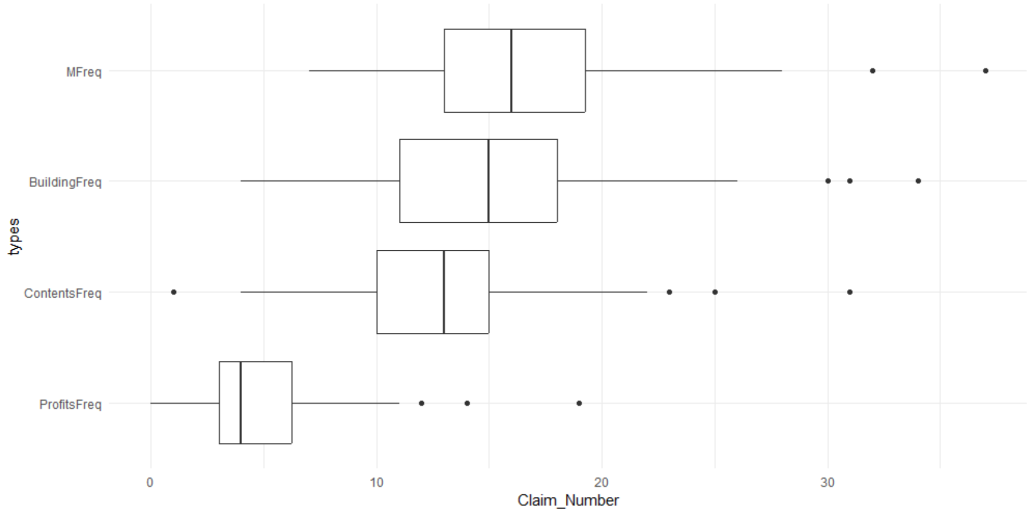
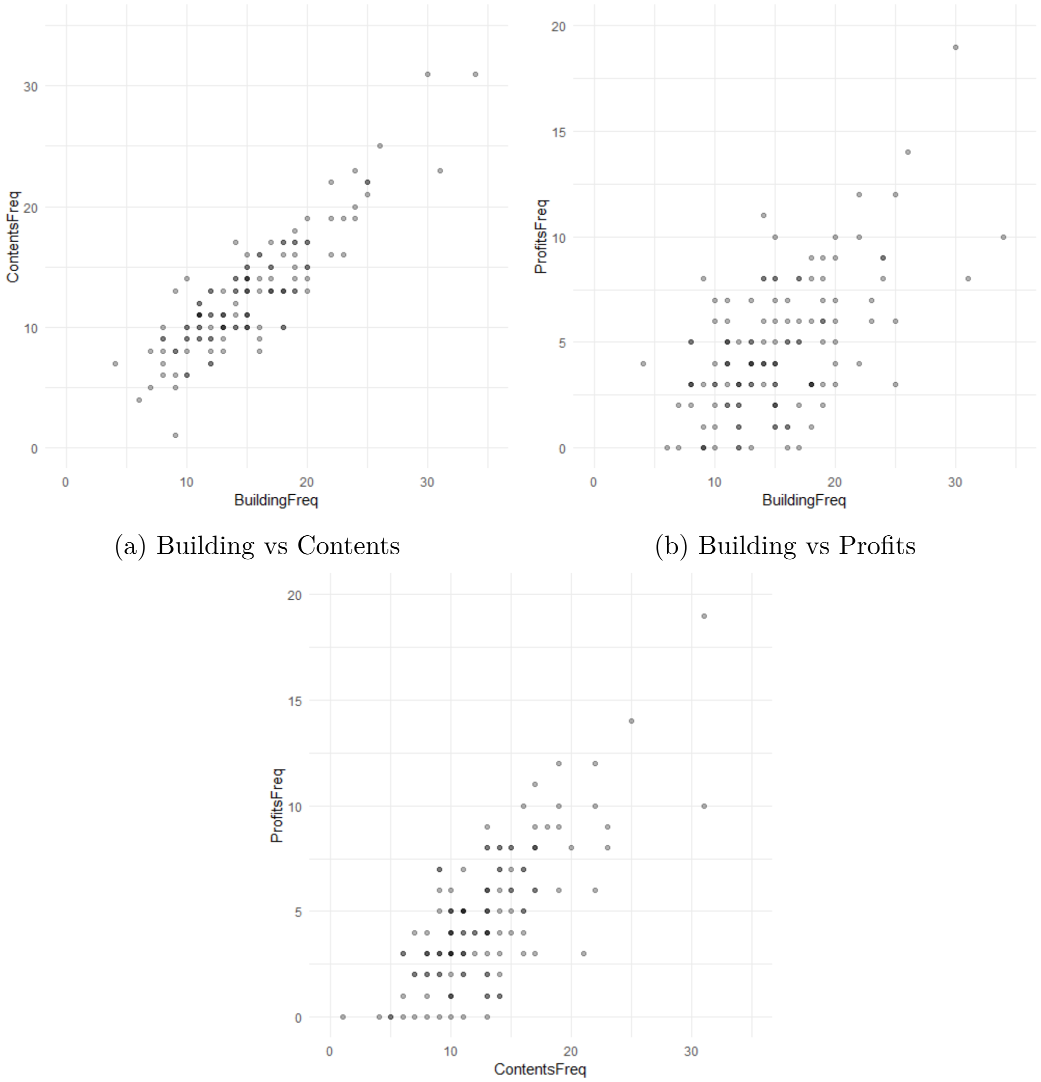


Figure 3.1: Claim Numbers' Boxplots

The claim numbers from three coverages are related based on the data structure and the intuitions. When checked using the Pearson correlation coefficient, the Building and Contents are highly correlated, with a coefficient of 0.88. Although the other two pairs are not larger than 0.88, they are still moderately correlated, where the Pearson correlation coefficients for Contents and Profits, and Building and Profits are 0.75 and 0.57, respectively. In addition, both Kendall and Spearman's rank correlation coefficients show positive relationships among the claims. Figure 3.2 shows the scatterplots in which we can visually observe the dependency between any two insurance coverages for the claim numbers.



(a) Building vs Contents

(b) Building vs Profits

(c) Contents vs Profits

Figure 3.2: Scatterplots of Three Pairs of Claim Numbers

In addition to our previous observations, we check whether the independent assumption is valid and whether there are seasonal effects. The lag 1-month autocorrelation statistic for the building claims is 0.21, which shows a weak time dependence. The profit sector has the most significant lag 1-month autocorrelation of 0.51. To observe seasonal effects, Figure 3.3 shows the boxplots of the claim numbers under the three insurance lines for each month of the year. The plot shows mild seasonal effects. On average, the claim numbers in summer and winter are observed to be greater than those in the other two seasons. In summary, although the observations show some time dependencies and mild seasonal effects, we assume time independence and focus on the dependency among coverages.

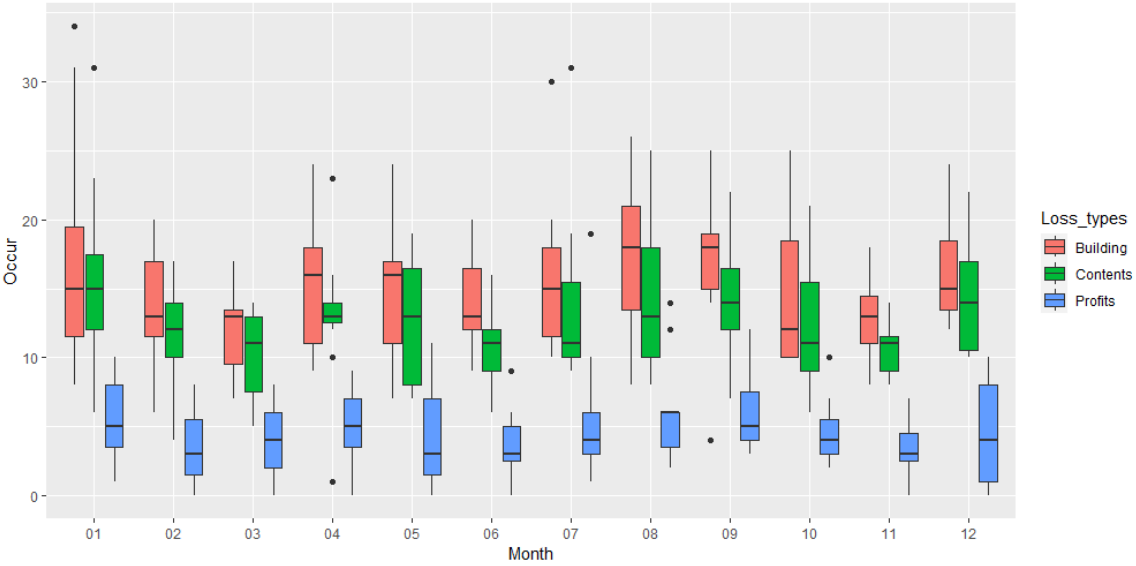


Figure 3.3: Seasonal Effects

Following the exploration of the frequency component, we further explore the individual losses after eliminating the zero losses. Table 3.3 displays the summary statistics regarding the individual losses for all three insurance coverages. From the table, we find that the loss amount under the building coverage has the largest summary statistics. We also observe from these statistics that all losses are heavy-tailed distributed. More specifically, we see that the mean is larger than the median for each of the three lines, and the maximum is quite significant compared to their corresponding third quartile statistics.

Source	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Building	0.02319	0.96618	1.32013	1.98668	1.97860	152.41321
Contents	0.00083	0.29000	0.57570	1.70178	1.44648	132.01320
Profits	0.00408	0.10011	0.26619	0.85180	0.67929	61.93265

Table 3.3: Summary of Loss Amount for Three Business Lines

Similar to the frequency data exploration, we show boxplots for losses in Figure 3.4. The boxplots demonstrate visually the heavy-tailed nature of the loss amounts and that the building coverage incurs the largest loss amounts on a per-claim basis. Note that Figure 3.4

shows only losses in $(0, 20]$; some extreme values are not presented in order to clearly show the right-skewed behaviour and detailed empirical distributions.

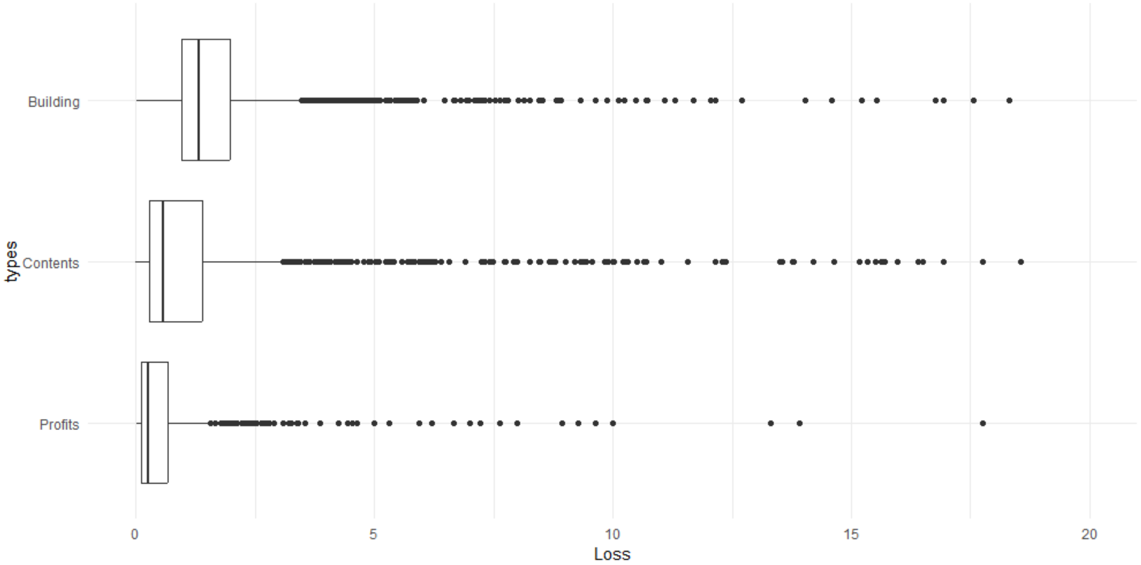


Figure 3.4: Individual Losses' Boxplots

3.2 Estimation Results

This section shows the estimation results for our proposed frequency and severity models. We first consider our benchmark model for the claim counts. Recall that for this model, the numbers of claims from the three insurance lines are all independently following negative binomial distributions. Specifically, $N_{it} \sim \mathcal{NB}(r_i, \lambda_i)$ where $t = 1, 2, \ldots, 132$, $i = 1, 2, 3$. For estimating the MLEs, we maximize the log-likelihood function given by (2.5) for all three coverages. Although the number of claims reported is irrelevant to this model, we apply a similar procedure to estimate its negative binomial parameters. The full log-likelihood function value is calculated by including the number of claims reported. However, there is no closed-form formula to calculate the MLE of the size parameter of the negative binomial distributions. To solve this, we apply the "optim" function in R to numerically obtain the estimated parameter values. Table 3.4 (right panel) shows the estimated values, the corresponding standard errors, and 95% interval estimates for this model. The standard errors of the estimates are generated from the "optim" function.

Unlike the Fully Independent model, the Binomial Thinning model requires working with full likelihood to estimate the parameter values because of the dependency between the claim reported and claims from three coverages. Recall that the marginal distribution of the claim reported follows a negative binomial with parameter r and λ, and the claim number for the i^{th} coverage $N_{it} \sim \mathcal{NB}(r, \lambda_i)$, where $t = 1, 2, \ldots, 132$. The logarithm of

Equation (2.7) is used to find the MLEs of the parameters for the Binomial Thinning model numerically. The left panel of Table 3.4 shows the estimated values and other relevant statistics.

	Binomial Thinning Model				Fully Independent Model			
	estimate	CI (95%)		SE	estimate	CI (95%)		SE
λ_1	15.08	14.24	15.91	0.43	15.08	14.21	15.95	0.44
r_1	-	-	-	-	20.74	8.96	32.52	6.01
λ_2	12.72	11.97	13.47	0.38	12.72	11.92	13.52	0.41
r_2	-	-	-	-	17.59	7.55	27.64	5.12
λ_3	4.67	4.27	5.07	0.20	4.67	4.11	5.22	0.28
r_3	-	-	-	-	3.62	1.94	5.30	0.86
λ	16.42	15.53	17.30	0.45	16.42	15.53	17.30	0.45
r	25.32	10.03	40.62	7.80	25.24	10.05	40.43	7.75
$\log \mathcal{L}$	-1183.47				-1516.57			
AIC	2382.94				3043.14			
BIC	2406.00				3057.56			

Table 3.4: Parameter estimates for the frequency components

From Table 3.4, we observe that the estimates for the λ values between the two models are similar (the same decimal places after rounding) as expected since λs represent the means in both models. For the building coverage, we obtain that the MLE of λ_1 is 15.08, which is the same as the λ_1 estimate under the Binomial Thinning model after rounding. Also, their corresponding standard errors are relatively small. The largest standard error of these λ estimates is 0.45 for the profits coverage. However, the standard errors of r estimates are quite large compared to those for all the λ estimates. While the standard error of the r_3 estimate under the Fully Independent model is 0.86, the standard errors of other r estimates are greater than 5, implying much wider interval estimates. For example, the 95% interval estimates of r under the Fully Independent model is $[10.05, 40.43]$. Besides these statistics, we obtained model comparison criteria for both. Recall from the model construction aspect that the Fully Independent model allows the situation that the number of claims reported is smaller than the claim number of single coverage. The Binomial Thinning model should outperform the Fully Independent model. AIC and BIC also suggest that the Binomial Thinning model is better than the independent one.

As mentioned in the previous chapter, the MLEs for the independent model can be used as estimators of the margins in Copula-Based models. Such parameter estimations maximize the profile likelihood instead of the complete likelihood by setting the copula parameter as a nuisance parameter. Then, by plugging in the estimated parameters of the margins, we maximize the log-likelihood function, (2.13), to estimate the copula parameter. The copula parameter estimation results are displayed in Table 3.5.

As mentioned previously, the parameter value for the Gaussian copula can be interpreted as a pairwise correlation level. For simplicity, we assume an exchangeable Gaussian copula where the correlation of all pairs is the same. Denoting the common correlation coefficient

	Gaussian Copula	Gumbel Copula	Joe Copula
Est. parameter	0.70452	1.83147	2.17170
Log-likelihood	-1015.953	-1021.079	-1033.461

Table 3.5: The estimates and log-likelihood of copula models

as ρ, the correlation matrix for the Gaussian copula can be expressed as

$$P = \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix}.$$

Our estimated Gaussian copula parameter is $\rho = 0.7$, which shows a strong positive correlation among the margins. The estimated parameters of the Gumbel and Joe copulas also suggest a positive and relatively strong dependency.

The composite models mentioned in Chapter 2 are fitted to the data, and the parameters are estimated. For applying composite models for the Building loss data, we consider the Exponential (E) and Gamma (G) distributions for the head part and the Inverse-Gamma (IG), Pareto (Pa), and Log-Normal (LN) distributions for the tail part of the losses. We then fit the six different composite models to the Building loss data. Table 3.6 shows the model selection criteria of these models. The model selection criteria for the losses from the other two business lines can be viewed in Appendix A.1

	G & IG	G & Pa	G & LN	E & IG	E & Pa	E & LN
# of parameters	4	4	4	3	3	3
$\log \mathcal{L}$	-2800.93	-2771.15	-2771.14	-3181.33	-3220.69	-3220.72
AIC	5609.87	5550.30	5550.29	6368.65	6447.37	6447.43
BIC	5632.25	5572.68	5572.67	6385.44	6464.16	6464.22

Table 3.6: Log-likelihood of composite models for the building losses

From the results in Table 3.6, the Gamma & Log-Normal (G & LN) composite and Gamma & Pareto (G & Pa) composite, both having four parameters, show similar fitting performance. Both models have almost the same log-likelihood values, as well as AIC and BIC values. We choose the Gamma & Log-Normal (G & LN) composite for our further analyses. Similar comparisons have been made for the contents and profits losses. We found that Gamma & Log-Normal (G & LN) composite model is the best for modelling content losses, and the Gamma & Pareto (G & Pa) fits the profit losses the best.

Table 3.7 shows the parameter estimates of three composite distributions, which are the best fit mentioned above. Using the density of composite distribution given by (2.15) for constructing the likelihood function, we estimate the parameters for the head and tail distributions. Thus, the splicing point and weight parameter can be obtained by using the expressions in Table 2.1 and Equation (2.17). Figure 3.5 shows the probability density functions of the three coverages' individual losses. We use different colors to indicate the head and tail part of the distribution.

	Building: G&LN	Contents: G&LN	Profits: G&Pa
Head Distribution	$\alpha_1 = 3.71085$ $\theta_1 = 0.37198$	$\alpha_1 = 1.98766$ $\theta_1 = 0.21591$	$\alpha_1 = 1.55072$ $\theta_1 = 0.10144$
Tail Distribution	$\mu_2 = -331.88884$ $\sigma_2 = 13.20987$	$\mu_2 = -1.34871$ $\sigma_2 = 1.69228$	$\alpha_2 = 1.41237$ $\theta_2 = 0.37195$
u	2.08943	0.47466	0.11282
ϕ	0.32151	1.34244	2.92302

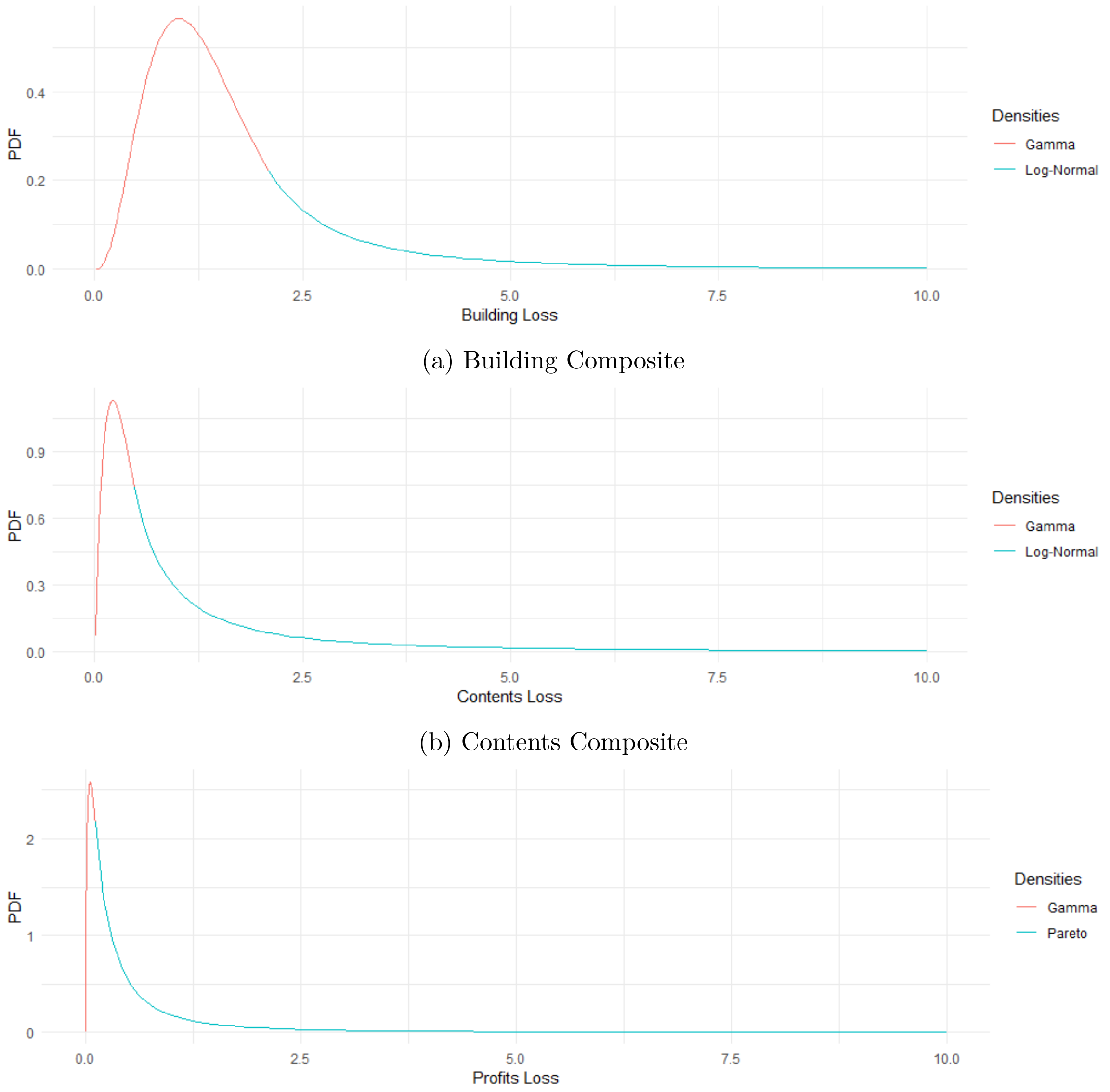
Table 3.7: Parameter estimates for the severity components

For interpreting the estimation results, we use the Gamma & Log-Normal for the building losses as an example. The parameter estimates for the head distribution are $\alpha_1 = 3.71085$ and $\theta_1 = 0.37198$. The estimates for the tail distribution, Log-Normal, are $\mu_2 = -331.88884$ and $\sigma_2 = 13.20987$. The splicing point for the composite distribution is calculated as $u = 2.08943$, which means that the whole distribution is split at 2.08943; the head part is Gamma distributed, and the tail part is Log-Normal distributed. Based on the estimated weight parameter, we can interpret $1/(1+\phi) = 0.24329$ as the proportion of Y that follows the Gamma distribution, and $\phi/(1+\phi) = 0.75671$ of Y follows the Log-Normal distribution.

In conclusion, we consider three main types of models: the Fully Independent model, the Binomial Thinning model, and three Copula-Based models. The Fully Independent model uses independent negative binomial random variables for the claim numbers from three insurance coverages, Building, Contents, and Profits, and their individual loss amounts are modelled by various composite models, where both Gamma & Log-Normal composite and Gamma & Pareto composite models are found outperformed. For the frequency component, the Binomial Thinning uses a negative binomial for the number of claims reported, and conditionally, the other three claim numbers are binomially distributed. The Copula-Based models link negative binomial claim numbers with Gaussian, Gumbel, and Joe copula structures. Recall that, the aggregate losses for the three types of model can be expressed by Equation (2.1).

3.3 Empirical Findings for Risk Management

In the previous section, we analyzed and compared the models from a statistical point of view. In this section, we study the models from an insurance perspective. Our primary interest is the performance of the risk analyses on the aggregation level for a specific time period (one month in this project), which aggregates the losses for all three insurance coverages. We conduct risk analyses based on the models that fit the data best in this chapter. Then, we assess how well the models performed in the insurance aspect.



(a) Building Composite

(b) Contents Composite

(c) Profits Composite

Figure 3.5: Composite Probability Density Functions for Three Coverages' Individual Losses

In the insurance industry, estimating the risk level for a product or portfolio is critical for determining the premium and reserve. Companies set the premium and reserve levels by considering the risks they face. Some common questions are, for example, how much the company should charge to make the produce profitable and still be competitive with the other companies, and how much capital the company should set aside to protect the company from bankruptcy with a 95% probability.

We recall some existing risk measures that we apply in this project. Let S be an aggregate loss random variable with cumulative distribution function F_S.

1. Value at Risk: $\mathrm{VaR}_\alpha(S) = \inf\{s \in \mathbb{R} : F_S(s) > \alpha\}$, $0 \leq \alpha \leq 1$, which can be treated as a standard to check the assets should be reserved to reduce the bankruptcy possibility to $1 - \alpha$.

2. Tail Value at Risk: $\mathbb{E}[S|S \geq \mathrm{VaR}_\alpha(S)]$, $0 \leq \alpha \leq 1$, which is the expected loss given that the losses exceeding the Value at Risk at α. That is, if bankruptcy happens, what the average excess loss is.

3. Proportional Hazard risk measure: $\mathrm{PH}_\alpha(S) = \int_0^\infty (1 - F_S(s))^{1/\alpha}\, ds,\ \alpha \geq 1$, introduced by Wang (1995). Specifically, the survival function is distorted in the definition. After the transformation, more probability is assigned to the extreme losses. Using the integral to calculate the expectation of the distorted random variable, we get a more conservative expected value.

4. Dual Power Risk Measure: $\mathrm{DP}_\beta(S) = \int_0^\infty 1 - (F_S(s))^\beta\, ds,\ \beta \geq 1$. The idea is similar to the Proportional Hazard risk measure except that the cumulative distribution function is now distorted.

As we can see, these risk measures all have their particular interpretation and advantages/disadvantages when using them for risk management. They should not be naively applied to all scenarios. For example, the Value at Risk provides information about the loss that will make the company insolvent. However, by fixing the α, the Value at Risk of an aggregate loss may be larger than the sum of the Value at Risks of each individual loss, which does not obey the idea of diversification. Artzner et al. (1999) proposed four criteria, and a risk measure satisfying them is called a coherent risk measure. The Tail Value at Risk is a coherent risk measure for continuous random variables. Wang (1994) proved the integration of the transformed distribution is coherent when the transformation is a concave function.

Because the closed-form distribution of the aggregate loss cannot be derived, we use the Monte Carlo simulation approach to approximate the risk measurements. We simulate 100,000 months of data, each including the reported claims, the claim numbers for three business lines, and the individual loss amounts for all three types of models. Further, the aggregate loss is the sum of individual losses for a month using (2.1).

The simulation process for the independent model is straightforward. We use random generations for the negative binomial to simulate claim numbers for all business lines. The Binomial Thinning model, on the other hand, requires the simulation of the reported claim numbers. We then apply binomial random generation with size parameters as the reported claim numbers to get the claim numbers for three lines of business. The simulation under the copula models uses a similar approach. We first randomly generate the uniform margins from the copula functions. With these uniform margins, we can obtain the claim numbers using the inverse of the marginal distributions. To simulate loss amounts from the composite models, we randomly generate the values from the standard uniform distribution and use the inverse of the composite distribution (2.16) to get simulated losses.

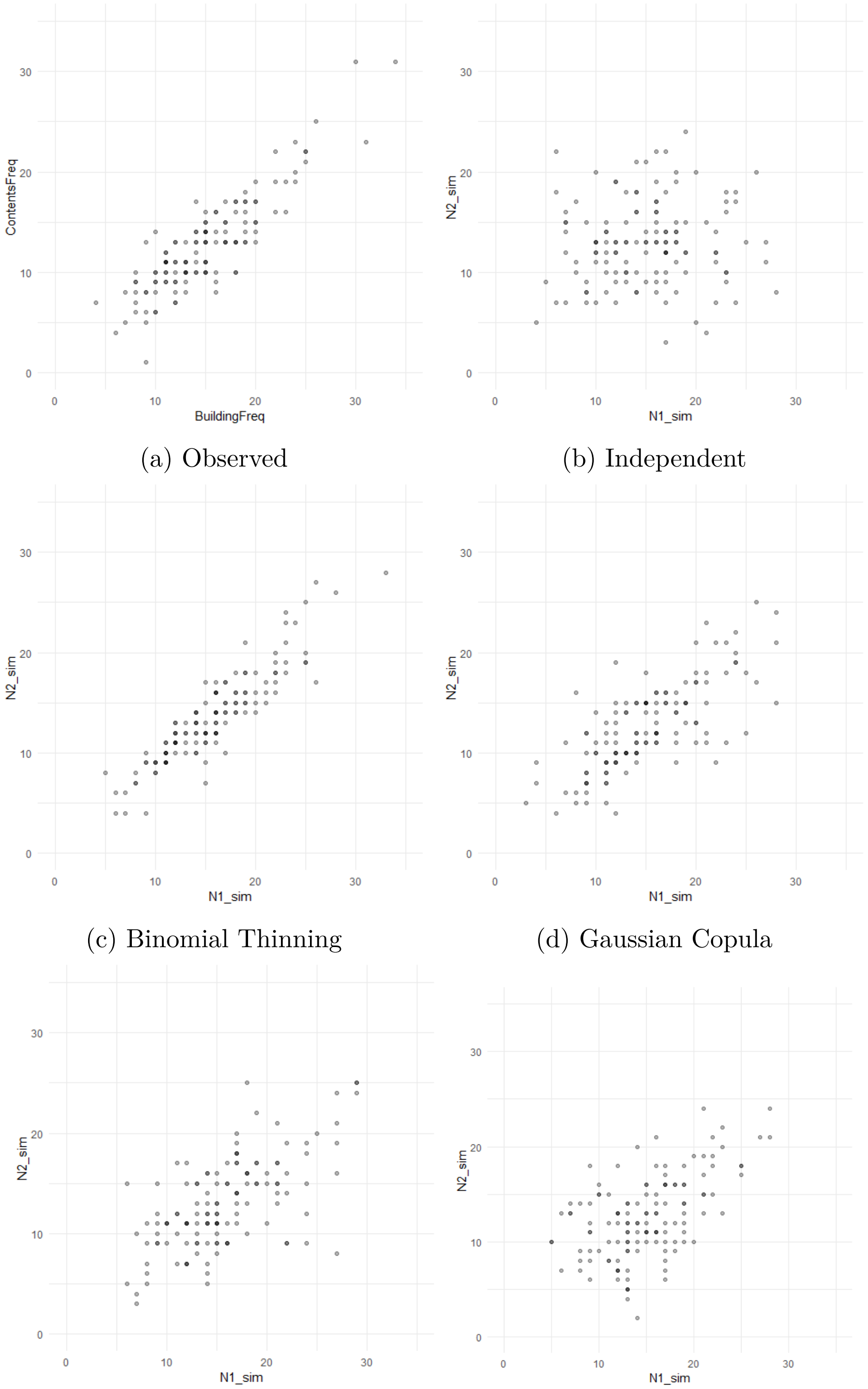
Figures 3.6 and 3.7 show the scatterplots of Building versus Contents and Building versus Profits claims for observed and simulated frequency data under the Fully Independent, the Binomial Thinning, and the three Copula-Based models. The scatterplot of Contents versus Profits can be viewed in Appendix A.2. Since the number of observations is 132 compared to 100000 months of simulated data points, we arbitrarily use 132 rows from the simulated data to make the scatterplots to easier compare with observed data.

Based on the plots of observed data in both Figures 3.6a and 3.7a, there are apparent positive relationships among the margins. As we expect, the independent model cannot capture such dependent behaviour. The Binomial Thinning model shows a strong positive dependency between the Building and Contents claim numbers, which is the closest one to the observed data visually. However, in Figure 3.7, the Joe copula appears to capture the most positive relationship between the Building and Profits. In addition, as we can see different copulas show different dependence structures. Gaussian copula has both the large claim numbers and the small claim numbers correlated closely in Figure 3.6d. In Figure 3.6f, Joe copula shows more correlations of the large claim numbers.

For each risk measure and each aggregate model, we investigate the risk measure for monthly aggregate losses for each insurance coverage and monthly total losses for the company. Table 3.8 shows the approximated risk measures under different aggregate models, in which Building, Contents, and Profits indicate the insurance coverages. The Aggregate column indicates the risk measures of the monthly total losses by aggregating the losses from three insurance coverages.

From Table 3.8, we can observe that the Fully Independent model underestimates most of its measures of risk for the Aggregate compared to the observed data. Especially, for the Tail Value at Risks, at $\alpha = 0.9$, the Fully Independent model suggests 114.7784 compared to the empirical 130.9614. The difference is magnified at $\alpha = 0.95$, where the estimate under the independent is 140.1614, while the empirical one is 171.8545. Both the Binomial Thinning and Copula-Based models perform better than the Fully Independent model and provide generally closer and risk measures comparable to the empirical estimates. Although the Tail Value at Risk at $\alpha = 0.95$ for the Binomial Thinning and Copula-Based models are all around 150, which still differ by around 20 compared to the empirical, they can be explained by the small sample size (we observed 132 months of data).

One finding regarding the Binomial Thinning and Copula-Based model is that the risk measures are similar except for the Proportional Hazard risk measures. The Gaussian copula's approximation is 134.76, which is the most conservative one compared to others. The Binomial Thinning and Joe copula provide relatively small and closer to the empirical results. It is hard to explain intuitively because proportional hazard risk measures distort the distributions of the aggregate loss for all models.



(a) Observed

(b) Independent

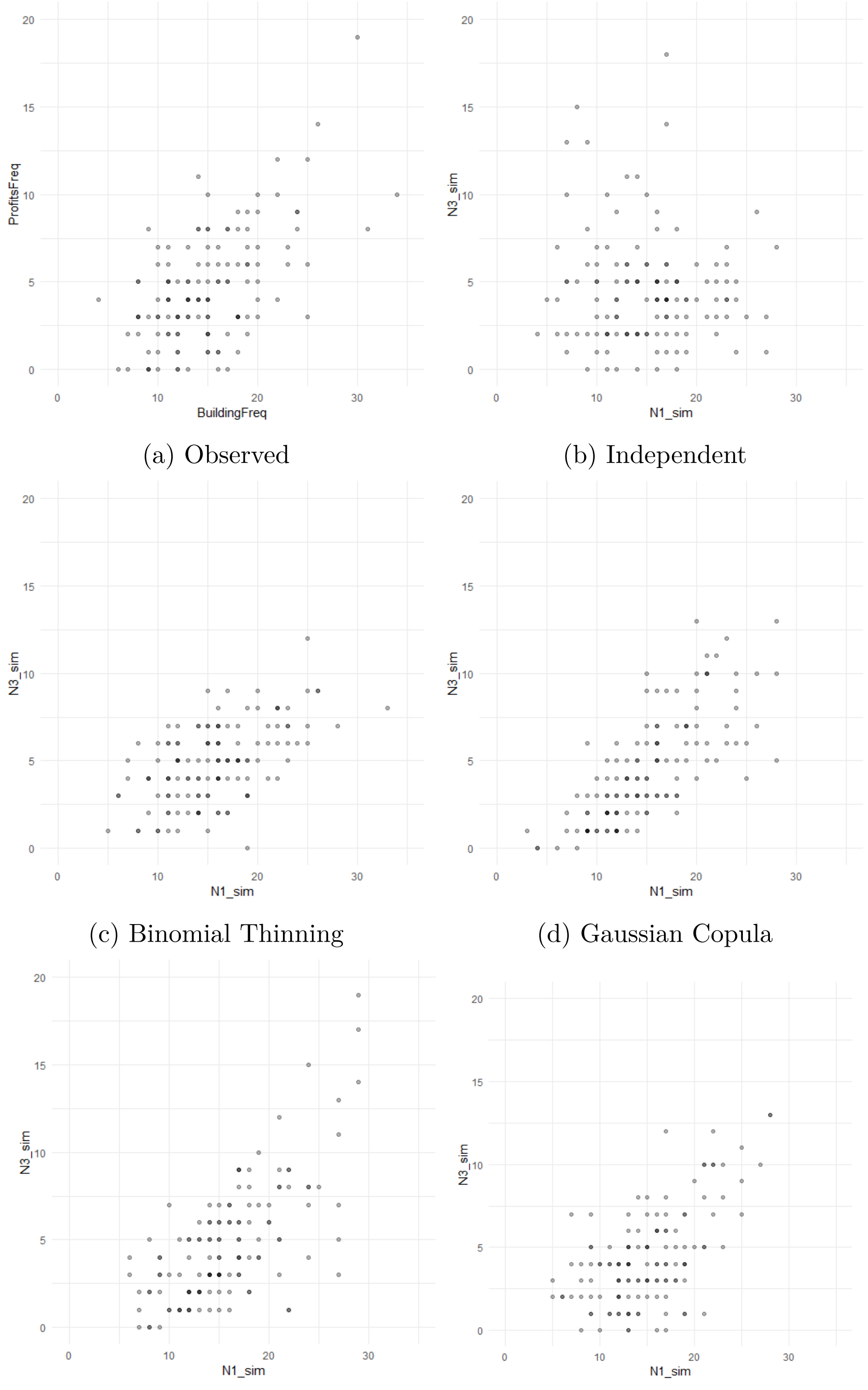
(c) Binomial Thinning

(d) Gaussian Copula

(e) Gumbel Copula

(f) Joe Copula

Figure 3.6: Observed & Simulated Building versus Contents Claims



(a) Observed

(b) Independent

(c) Binomial Thinning

(d) Gaussian Copula

(e) Gumbel Copula

(f) Joe Copula

Figure 3.7: Observed & Simulated Building versus Profits Claims

Measure	Model	Building	Contents	Profits	Aggregate
$\mathrm{VaR}_{0.90}(S)$	Observations	43.3675	37.6107	8.92768	84.9583
	Independent	45.8395	40.1350	8.8537	82.5497
	Binomial Thinning	45.5542	39.8117	8.3302	87.4485
	Gaussian	46.0096	39.9315	8.9250	88.0304
	Gumbel	45.9438	40.4155	8.8096	88.6401
	Joe	46.0086	40.2537	8.7924	88.3784
$\mathrm{TVaR}_{0.90}(S)$	Observations	67.6329	61.7309	17.3499	130.9614
	Independent	62.2931	64.1973	21.7799	114.7784
	Binomial Thinning	62.1844	63.4675	22.5703	122.2398
	Gaussian	62.5756	63.7182	24.1390	123.6911
	Gumbel	62.7209	64.0983	24.1995	124.4939
	Joe	63.2035	63.4808	22.3663	122.8164
$\mathrm{TVaR}_{0.95}(S)$	Observations	88.5959	82.6582	24.1481	171.8545
	Independent	75.1908	82.9212	32.7698	140.1614
	Binomial Thinning	75.2890	82.0129	34.9973	149.5173
	Gaussian	75.5330	82.1879	37.4004	151.7549
	Gumbel	75.8074	82.2438	37.6642	152.5675
	Joe	76.6455	81.3573	34.0908	149.5723
$\mathrm{PH}_2(S)$	Observations	51.9328	42.8176	11.3142	93.6248
	Independent	54.1955	56.3798	28.3590	102.4392
	Binomial Thinning	58.8418	56.8998	40.0197	114.9348
	Gaussian	53.3108	50.3601	63.6347	134.7637
	Gumbel	53.5260	49.7185	55.6399	126.7038
	Joe	60.1957	47.3558	38.5885	111.7715
$\mathrm{DP}_3(S)$	Observations	43.2273	35.9657	8.2477	82.2996
	Independent	41.5724	35.7542	9.2529	76.3109
	Binomial Thinning	41.5093	35.5262	9.4621	79.5965
	Gaussian	41.6905	35.6027	9.9802	80.0948
	Gumbel	41.6977	35.8246	9.9892	80.2022
	Joe	41.8295	35.5719	9.4253	79.6027

Table 3.8: Approximated Risk Measures

Chapter 4

Conclusion and Discussion

Insurance companies need to manage risks. They accept risks transferred from individuals or companies in exchange for premiums. Almost every step in the operation of an insurance company involves risk management, including setting profitable and competitive premium levels and managing the capital to cover potential losses.

However, risk should be understood correctly. Statistical models should closely reflect the empirical behaviours to provide informative suggestions. In this project, we found capturing the dependency among different insurance lines and the heavy-tailed behaviour for the per-claim losses can improve the performance of assessing the risk of monthly aggregate loss.

Besides, the dependencies among claims from the dataset behave the closest to the Binomial Thinning dependencies, which are all linear dependent. In this case, the Binomial Thinning is already the best fit for the data compared to other Copula-Based models. However, we still find the Copula-Based model able to combine different dependencies with discrete random variables. Further, unlike the Hierarchical models, the construction of the Copula-Based model does not require intuition understanding regarding the data itself.

For further research, Geenens (2020) proposed a way to connect discrete random variables with a dependent structure. The logic is similar to the copula but with more mathematical rigours.

Bibliography

P. Artzner, F. Delbaen, E. Jean-Marc, and D. Heath. Coherent measures of risk. *Mathematical Finance*, 9:203 – 228, 07 1999. doi: 10.1111/1467-9965.00068.

A. Bücher and C. Zhou. A horse race between the block maxima method and the peak–over–threshold approach. *Statistical science*, 36(3):360, 2021. ISSN 0883-4237.

S. Cabras and M. E. Castellanos. A bayesian approach for estimating extreme quantiles under a semiparametric mixture model. *ASTIN Bulletin*, 41(1):87–106, 2011.

A. C. Cameron, T. Li, P. K. Trivedi, and D. M. Zimmer. Modeling the differences in counted outcomes using bivariate copula models: with application to mismeasured counts. *Econometrics Journal*, 7:566–584, 12 2004. doi: 10.1111/j.1368-423X.2004.00144.x.

D. G. Clayton. 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, 1978.

K. Cooray and M. M. Ananda. Modeling actuarial data with a composite lognormal-Pareto model. *Scandinavian Actuarial Journal*, 2005(5):321–334, 2005.

M. Eling. Fitting insurance claims to skewed distributions: Are the skew-normal and skew-student good models? *Insurance: Mathematics and Economics*, 51(2):239–248, 2012.

R. A. Fisher and L. H. C. Tippett. Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society*, 24(2):180–190, 1928. doi: 10.1017/S0305004100015681.

T. C. Fung, H. Jeong, and G. Tzougas. Investigating the effect of climate-related hazards on claim frequency prediction in motor insurance. *Available at SSRN 4638074*, 2023.

T. C. Fung, H. Jeong, and G. Tzougas. Soft splicing model: bridging the gap between composite model and finite mixture model. 2024(2):168–197, 2024. ISSN 0346-1238.

G. Geenens. Copula modeling for discrete random vectors. *Dependence Modeling*, 8:417–440, 12 2020. doi: 10.1515/demo-2020-0022.

B. Gnedenko. Sur la distribution limite du terme maximum d'une série aléatoire. *Annals of Mathematics*, 44(3):423–453, 1943. doi: 10.2307/1968974.

E. J. Gumbel. Distributions des valeurs extremes en plusieus dimensions. *Publications de l'Institut de Statistique de l'Université de Paris*, 9:171–173, 1960.

L. Hong and R. Martin. Dirichlet process mixture models for insurance loss data. *Scandinavian Actuarial Journal*, 2018(6):545–554, 2018.

H.-M. Hou and Y. Liu. Analysis of the upper tail of the short-term extreme tension distribution of mooring line by the peaks-over-threshold method. *Ocean engineering*, 281: 114994, 2023. ISSN 0029-8018.

H. Jeong, G. Tzougas, and T. C. Fung. Multivariate claim count regression model with varying dispersion and dependence parameters. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 186(1):61–83, 2023.

H. Joe. Parametric families of multivariate distributions with given margins. *Journal of Multivariate Analysis*, 46:262–282, 08 1993. doi: 10.1006/jmva.1993.1061.

A. Lee. Modelling rugby league data via bivariate negative binomial regression. *Australian New Zealand Journal of Statistics*, 41:141 – 152, 12 2002. doi: 10.1111/1467-842X.00070.

A. J. McNeil. Estimating the tails of loss severity distributions using extreme value theory. *ASTIN Bulletin*, 27(1):117–137, 8 1997. doi: 10.2143/AST.27.1.563210.

T. Miljkovic and B. Grün. Modeling loss data using mixtures of distributions. *Insurance: Mathematics and Economics*, 70:387–396, 2016.

A. K. Nikoloulopoulos. Copula-based models for multivariate discrete response data. pages 231–249, 2013. doi: 10.1007/978-3-642-35407-6_11.

R. Oh, H. Jeong, J. Y. Ahn, and E. A. Valdez. A multi-year microlevel collective risk model. *Insurance: Mathematics and Economics*, 100:309–328, 2021.

F. Pechon, M. Denuit, and J. Trufin. Home and motor insurance joined at a household level using multivariate credibility. *Annals of Actuarial Science*, 15(1):82–114, 2021.

M. Pigeon and M. Denuit. Composite lognormal–Pareto model with random threshold. *Scandinavian Actuarial Journal*, 2011(3):177–192, 2011.

S. I. Resnick. Discussion of the danish data on large fire insurance losses. *ASTIN bulletin*, 27(1):139–151, 1997. ISSN 0515-0361.

D. P. Scollnik and C. Sun. Modeling with Weibull-Pareto models. *North American Actuarial Journal*, 16(2):260–272, 2012.

A. Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut Statistique de l'Université de Paris*, (8):229–231, 1959.

R. Vernic, C. Bolancé, and R. Alemany. Sarmanov distribution for modeling dependence between the frequency and the average severity of insurance claims. *Insurance: Mathematics and Economics*, 102, 12 2021. doi: 10.1016/j.insmatheco.2021.12.001.

S. Wang. Premium calculation by transforming the layer premium density. *ASTIN Bulletin*, 26, 09 1994. doi: 10.2143/AST.26.1.563234.

S. Wang. Insurance pricing and increased limits ratemaking by proportional hazards transforms. *Insurance: Mathematics and Economics*, 17:43–54, 08 1995. doi: 10.1016/0167-6687(95)00010-P.

Appendix A

Additional Tables and Figures

A.1 Contents and Profits' Individual Losses Modelling Selection Results

We also considered six composite distributions for the losses of content and profit. From the following table, the Gamma & Log-Normal composite distribution best fits the content loss. The Gamma & Pareto composite distribution can model the profit loss.

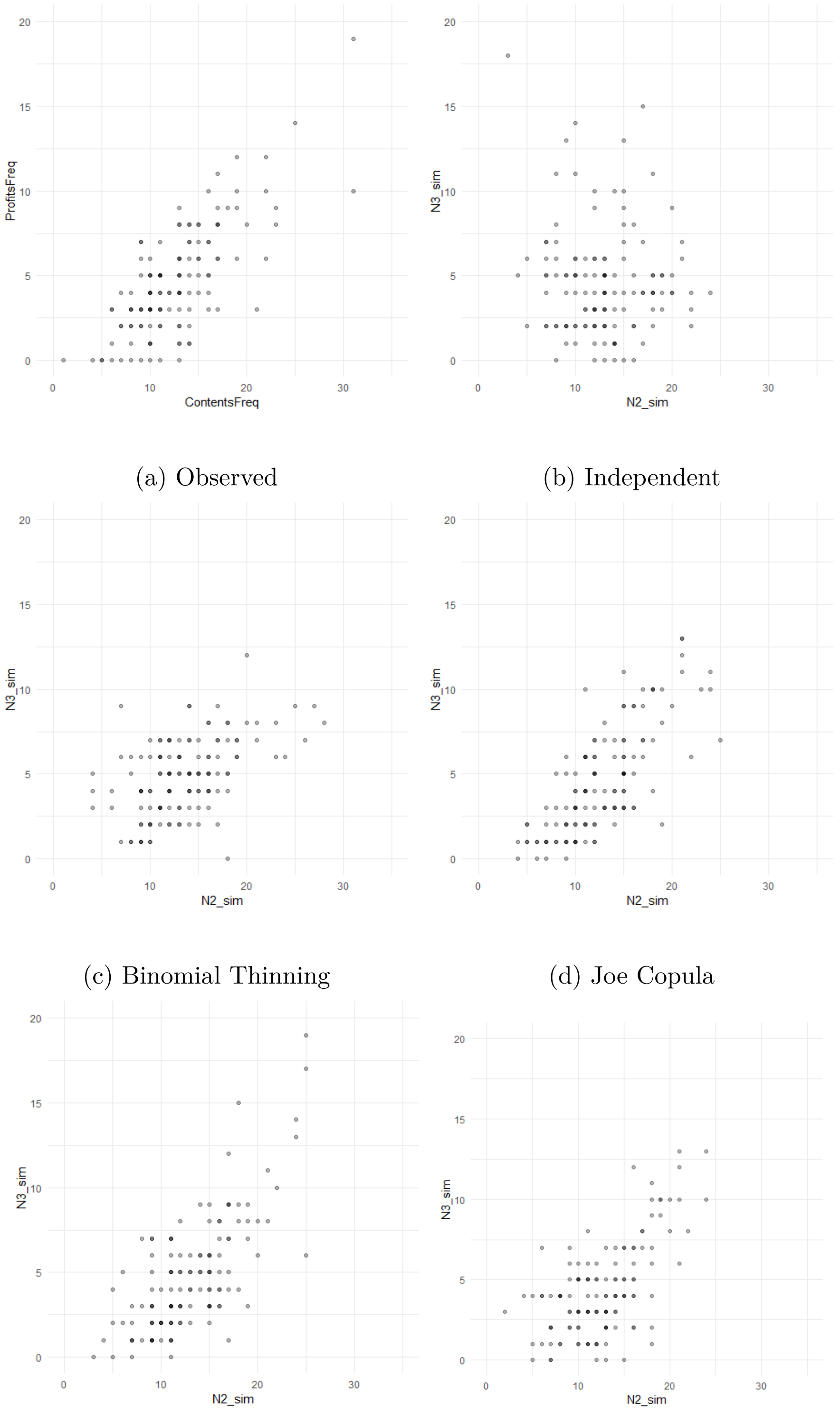
	G & IG	G & Pa	G & LN	Exp & IG	Exp & Pa	Exp & LN
# of parameters	4	4	4	3	3	3
$\log \mathcal{L}$	-2187.88	-2039.52	-2037.59	-2102.97	-2102.81	-2102.16
AIC	4383.77	4087.04	4083.18	4211.95	4211.61	4210.32
BIC	4405.47	4108.74	4104.88	4228.23	4227.89	4226.60

Table A.1: Log-likelihood of composite models for the content losses

	G & IG	G & Pa	G & LN	Exp & IG	Exp & Pa	Exp & LN
# of parameters	4	4	4	3	3	3
$\log \mathcal{L}$	-309.19	-297.19	-427.81	-305.93	-304.53	-304.48
AIC	626.38	602.39	863.62	617.86	615.06	614.97
BIC	644.07	620.08	881.31	631.13	628.33	628.24

Table A.2: Log-likelihood of composite models for the profit losses

A.2 Observed and Simulated Contents vs Profits Claims



(a) Observed

(b) Independent

(c) Binomial Thinning

(d) Joe Copula

(e) Joe Copula

(f) Joe Copula

Figure A.1: Observed & Simulated Building vs Profits Claims

Appendix B

Code

The R codes are available at `https://github.com/AnxiousLegHair/TianxingYan_masterproject.git`.