

Article

A Copula-Based Bivariate Composite Model for Modelling Claim Costs

Girish Aradhye^{1,†}, George Tzougas²  and Deepesh Bhati^{1,*,†} ¹ Department of Statistics, Central University of Rajasthan, Ajmer 305817, India; 2020phdsta003@curaj.ac.in² Maxwell Institute for Mathematical Sciences, Department of Actuarial Mathematics and Statistics, Heriot-Watt University, Edinburgh EH14 4AS, UK; george.tzougas@hw.ac.uk

* Correspondence: deepesh.bhati@curaj.ac.in

† These authors contributed equally to this work.

Abstract: This paper aims to develop a new family of bivariate distributions for modelling different types of claims and their associated costs jointly in a flexible manner. The proposed bivariate distributions can be viewed as a continuous copula distribution paired with two marginals based on composite distributions. For expository purposes, the details of one of the proposed bivariate composite distributions is provided. The dependence measures for the resulting bivariate copula-based composite distribution are studied, and its fitting is compared with other bivariate composite distributions and existing bivariate distributions. The parameters of the proposed bivariate composite model are estimated via the inference functions for margins (IFM) method. The suitability of the proposed bivariate distribution is examined using two real-world insurance datasets, namely the motor third-party liability (MTPL) insurance dataset and Danish fire insurance dataset.

Keywords: copulas; dependence parameter; Gumbel copula; Inverse Weibull distribution; Inverse Burr distribution; Paralogistic distribution; Weibull distribution

MSC: 60E05; 62F10; 62P05; 91G70



Citation: Aradhye, G.; Tzougas, G.; Bhati, D. A Copula-Based Bivariate Composite Model for Modelling Claim Costs. *Mathematics* **2024**, *12*, 350. <https://doi.org/10.3390/math12020350>

Academic Editor: Luca Piccoli

Received: 16 December 2023

Revised: 16 January 2024

Accepted: 19 January 2024

Published: 22 January 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

During the last few decades, there has been a notable surge in actuarial research focused on modelling the costs of different types of claims in non-life insurance, employing a diverse range of claim severity approaches. This is due to the peculiar characteristics of the claim severity distribution, which pose several challenges. The distribution often spans several magnitudes, encompassing small and moderate claim sizes with high frequency, as well as a few major ones with low frequency. Additionally, claim size data are unimodal and heavily skewed to the right (see, for instance, Ref. [1]). As can be clearly understood, when the data span a wide range of magnitudes, selecting a probability distribution that can efficiently fit small, moderate, and large claims becomes crucial for insurance pricing, reserving, and risk management. The method of composing distributions (see, for example, Refs. [2–6]) provides a reasonable fit for such datasets.

Regarding the most recent studies on composite models, which are the main research focus of this work, it is worth noting that Ref. [6] proposed different composite models by considering Burr, Loglogistic, Paralogistic, and Generalized Pareto distributions for the tail of the data and truncated densities before and after the threshold point. Ref. [7], instead of creating single composite model, considered 256 composite models derived from sixteen parametric distributions frequently used in actuarial science. Ref. [8] placed special emphasis on modelling extreme claims using a variety of composite models and threshold selection techniques, including heuristic methods, the Minimum AMSE of the Hill estimator, the exponentiality test, and the Gertensgarbe plot. A new composite distribution known as the composite Rayleigh–Pareto (CRP) distribution was presented by [9]. This article also

discussed parameter estimation methods, including the method of moment and maximum likelihood estimation method. Ref. [10] introduced a novel single-parameter composite length-biased exponential Pareto (CLBEP) distribution to model insurance losses. The distributional properties and other statistical characteristics of the CLBEP distribution are mathematically determined. Authors provided a comparison study of the new CLBEP with other composite and conventional distributions.

Ref. [11] introduced a mixture composite claim severity regression model extending the setup of Ref. [12], which used a finite mixture distribution for the body and a Pareto-type distribution for the tail of the distribution. This extension incorporated explanatory variables on all three parts of the claim size distribution: clustering probabilities, body part, and tail part. It should be noted that, even though the literature concerning composite models in the univariate setting is abundant, their bivariate extensions have not been investigated so far. However, in non-life insurance, it is common for the actuary to observe the existence of dependence structures between different types of claims and their associated costs, either from the same type of coverage or from multiple types of coverage, such as motor and home insurance bundled into one single policy.

Examining interdependent hazards is a pivotal undertaking for insurance companies. The presence of dependence is manifested in the ability to predict the distribution of one random variable based on the knowledge of another. Particularly in the realm of insurance, it is imperative for companies to delve into the interconnections between various lines of business and assess the repercussions of a catastrophic event, such as an earthquake or hurricane, impacting multiple lines simultaneously. Copulas emerge as a prominent model-based approach for scrutinizing dependencies in risks, and the tail dependence coefficient serves as a metric for measuring dependence concerning extreme losses. The rising popularity of utilizing copulas to compute joint distributions for two or more random variables, depicting extreme losses, and capturing tail dependence is evident in the fields of actuarial science and finance Refs. ([13,14]). The copula technique is distinctive for partitioning the joint distribution into two components: individual marginal distributions for each random variable and a copula function that amalgamates the marginals to form a joint distribution explaining the dependence structure. Though direct modeling with multivariate distributions is generally preferred for joint risks, complexity in specification and limited applications can pose challenges. Copulas provide an alternative, allowing a focus on effective independent marginal risk modeling before uniting them using a copula framework. This approach gains prominence due to the wealth of literature available for univariate modeling, making it a practical choice in various scenarios.

Copula functions offer an efficient method to quantify dependence between hazards by fully specifying the dependence structure among random variables. Understanding tail dependence across diverse business lines is vital for insurance and reinsurance. Unlike multivariate distributions, copulas simplify summarizing tail dependence with diverse dependency structures achievable through various parameter settings. Archimedean copulas like Gumbel and Joe focus on modeling right tail extremes, Clayton on left tail extremes, and symmetric copulas like Frank, Gaussian, and t copulas distribute equal weight to both tails. Considering this fact, in the present study, we introduce a Gumbel copula-based bivariate distribution having composite marginal distributions for modelling heavy right-tailed claim severity jointly from multiple types of claims, which are often encountered in high-dimensional non-life insurance datasets. The motivation behind the development of the bivariate composite models mainly lies in the following factors:

1. Insurance portfolios often involve multiple risks that may be correlated. A bivariate distribution allows for modeling the joint behavior of two variables, capturing the dependence structure between them. This is crucial for accurately assessing the overall risk in a portfolio.
2. In insurance, the occurrence of extreme events is of particular interest. Bivariate composite distributions can better capture tail dependence, which is essential for estimating the joint probability of simultaneous extreme events.

Bivariate composite distributions are being developed in order to describe insurance heavy-tailed datasets. This is because more realistic and accurate representations of joint risk structures in complicated insurance portfolios are required. In particular, we make the following contributions:

1. The bivariate distributions are constructed using a suitable copula distribution consisting of two marginal composite distributions.
2. A general framework for constructing the copula-based bivariate composite distribution is discussed. For expository purposes, a specific bivariate composite distribution is presented.
3. This distribution aims to describe the behavior of bivariate data with a predominance of small and medium values but includes a few extreme values.
4. Dependence measures associated with the proposed bivariate composite distribution based on the copula distribution are derived.
5. The parameters of the distributions are estimated using the IFM method, which consists of estimating univariate parameters by separately maximizing the marginal composite distribution and then estimating the dependence parameters from the bivariate likelihoods derived based on the copula.

The rest of the paper is structured as follows. In Section 2, we derive alternative marginal composite distributions based on the Classical Composition (CC) technique. Section 3 presents the construction of the bivariate composite distributions based on the copula distribution. Parameter estimation of the bivariate composite distribution via the IFM method is discussed. The computational aspects of fitting the bivariate composite distributions are discussed in Section 4. Section 5 demonstrates a data generation algorithm and results based on a simulation experiment. In Section 6, we describe the MTPL and Danish fire loss datasets used for our empirical analysis and comparison of the bivariate composite distributions with existing bivariate distributions. Finally, concluding remarks can be found in Section 7.

2. Modelling Framework

2.1. The Univariate Composite Distribution Obtained through the Traditional Composition Method

Ref. [1] proposed various composite distributions using an unrestricted mixing weight (r), the right-truncated and left-truncated densities truncated at thresholds (θ) for the head and tail distributions, respectively. The resulting probability density function (pdf) of the composite distribution can be written as

$$f(x) = \begin{cases} rf_1^*(x|\Xi_1, \theta) & \text{for } 0 < x \leq \theta, \\ (1-r)f_2^*(x|\Xi_2, \theta) & \text{for } \theta < x < \infty, \end{cases} \quad (1)$$

where Ξ_1 and Ξ_2 are the parameter spaces for the head and tail parts of the composite distribution, and the mixing weight $r \in [0, 1]$ and $\theta > 0$. The functions $f_1^*(x|\Xi_1, \theta) = \frac{f_1(x|\Xi_1)}{F_1(\theta|\Xi_1)}$ and $f_2^*(x|\Xi_2, \theta) = \frac{f_2(x|\Xi_2)}{1-F_2(\theta|\Xi_2)}$ are the adequate truncations of the pdfs f_1 and f_2 up to and after an unknown threshold value θ , respectively.

- The value of the weight parameter r is obtained by the continuity condition imposed at the threshold θ , i.e., $rf_1^*(\theta|\Xi_1, \theta) = (1-r)f_2^*(\theta|\Xi_2, \theta)$. Hence, we obtain

$$r = r(\theta, \Xi_1, \Xi_2) = \frac{f_2(\theta|\Xi_2)F_1(\theta|\Xi_1)}{f_2(\theta|\Xi_2)F_1(\theta|\Xi_1) + f_1(\theta|\Xi_1)(1-F_2(\theta|\Xi_2))}. \quad (2)$$

- Further, imposing the differentiability condition at the threshold value θ , i.e., $rf_1^{*'}(\theta|\Xi_1, \theta) = (1-r)f_2^{*'}(\theta|\Xi_2, \theta)$, makes the density smooth.

These above conditions reduce the number of parameters and make the resulting density continuous and differentiable. We henceforth refer to this technique as the Classical Composition (CC) technique.

2.2. The Copula

A copula $C(u_1, u_2)$ is a bivariate cumulative distribution function (cdf) with uniform marginals (u_1, u_2) on the interval $(0, 1)$ Ref. ([15,16]). If $F_j(y_j)$ is the cdf of a univariate r.v. Y_j , then $C(F_1(y_i^{(1)}), F_2(y_i^{(2)}))$ is a bivariate distribution for $Y = (Y^{(1)}, Y^{(2)})$ with marginal distributions F_j , where $i = 1, 2, \dots, n$ and $j = 1, 2$.

The dependence of Y_i among the claim types is modelled using the copula, with the joint distribution of Y_i given by

$$\pi(y_i) = C(u_1, u_2) = C_\phi\left(F_1\left(y_i^{(1)}\right), F_2\left(y_i^{(2)}\right)\right), \quad (3)$$

where $y_i = (y_i^{(1)}, y_i^{(2)})$ are the realizations of the $Y = (Y^{(1)}, Y^{(2)})$, C_ϕ is a copula function, and $\phi := \{\phi^{(j,j')}\}_{j,j'=1,2}$ is the copula parameter that explains the association between the two random variables, say $(Y^{(1)}, Y^{(2)})$.

2.3. Modelling Dependence via the Gumbel Copula

Many insurance datasets exhibit strong correlation at high values of claim amounts but less correlation at low values of claim amounts. Hence, to jointly model the two types of claims having high tail dependence, the Gumbel copula will be an appropriate choice to model such a dataset (see Ref. [17]). The dependency between two types of claims $y_i^{(1)}, y_i^{(2)}$ can be modelled through the Gumbel copula as

$$C(u_1, u_2) = C_\phi\left(F_1(y_i^{(1)}), F_2(y_i^{(2)})\right) = \exp\left\{-\left[\left(-\log F_1(y_i^{(1)})\right)^\phi + \left(-\log F_2(y_i^{(2)})\right)^\phi\right]^{\frac{1}{\phi}}\right\}, \quad (4)$$

where $\phi := \{\phi^{(j,j')}\}_{j,j'=1,2} \in [1, \infty)$ is the asymmetric copula parameter influencing the correlations among $(Y^{(1)}, Y^{(2)})$. $F_j(y_j)$ is the cdf of a univariate r.v. Y_j for $j = 1, 2$. We present two dependence measures associated with the Gumbel copula as:

- Kendall's tau: Kendall's tau, denoted by τ , is a bivariate measure of dependence for continuous variables that measures the amount of concordance present in a bivariate distribution. Kendall's tau for the Gumbel copula in terms of the copula parameter ϕ can be written as:

$$\tau = \frac{\phi - 1}{\phi}.$$

- Tail dependence property: The amount of dependency in the upper or lower quadrant tail of a bivariate distribution is referred to as bivariate tail dependence. The expression for the upper tail dependence parameter λ_U for the Gumbel copula is given by:

$$\lambda_U = 2 - 2^{\frac{1}{\phi}}.$$

3. The Bivariate Composite H-Inverse Weibull Distribution

Let $Y = (Y^{(1)}, Y^{(2)})$ and $y_i = (y_i^{(1)}, y_i^{(2)})$, respectively, be the claims vector and its corresponding realizations of the two types of claims. Suppose that $Y_i^{(j)}$, $j = 1, 2$ follows the composite H-Inverse Weibull (IW) distribution, where H stands for the head part distribution of the bivariate composite distribution with pdf given by

$$f_j(y_i^{(j)}) = \begin{cases} r_{H,IW}^{(j)} f_H^*(y_i^{(j)}; \Xi^{(j)}) & \text{for } 0 < y_i^{(j)} \leq \theta^{(j)}, \\ (1 - r_{H,IW}^{(j)}) \frac{\frac{\alpha^{(j)}}{y_i^{(j)}} \left(\frac{\gamma^{(j)}}{y_i^{(j)}}\right)^{\alpha^{(j)}} \exp\left\{-\left(\frac{\gamma^{(j)}}{y_i^{(j)}}\right)^{\alpha^{(j)}}\right\}}{1 - \exp\left\{-\left(\frac{\gamma^{(j)}}{\theta^{(j)}}\right)^{\alpha^{(j)}}\right\}} & \text{for } \theta^{(j)} < y_i^{(j)} < \infty. \end{cases} \quad (5)$$

The cdf of composite H -Inverse Weibull distribution may be written as

$$F_j(y_i^{(j)}) = \begin{cases} r_{H,IW}^{(j)} \frac{F_H(y_i^{(j)}; \Xi^{(j)})}{F_H(\theta^{(j)}; \Xi^{(j)})} & \text{for } 0 < y_i^{(j)} \leq \theta^{(j)}, \\ r_{H,IW}^{(j)} + (1 - r_{H,IW}^{(j)}) \frac{\exp\left\{-\left(\frac{\gamma^{(j)}}{y_i^{(j)}}\right)^{\alpha^{(j)}}\right\} - \exp\left\{-\left(\frac{\gamma^{(j)}}{\theta^{(j)}}\right)^{\alpha^{(j)}}\right\}}{1 - \exp\left\{-\left(\frac{\gamma^{(j)}}{\theta^{(j)}}\right)^{\alpha^{(j)}}\right\}} & \text{for } \theta^{(j)} < y_i^{(j)} < \infty, \end{cases} \quad (6)$$

where $i = 1, 2, \dots, n$ and $j = 1, 2$. $\Xi^{(j)}$ and $r_{H,IW}^{(j)} \in [0, 1]$ are the parameter space and mixing weight for the head part of the j^{th} marginal composite H -Inverse Weibull distribution, $\alpha^{(j)} > 0$, threshold point associated with j^{th} marginal composite distribution $\theta^{(j)} > 0$, and scale parameter $\gamma^{(j)} > 0$. $F_H(y^{(j)})$ and $F_H(\theta^{(j)})$ are the cdf of the H (head) distribution at $y^{(j)}$ and the threshold point $\theta^{(j)}$, respectively. $f_H(\cdot)$ is the density of the head part of the marginal composite H -Inverse Weibull distribution. $f_H^*(y^{(j)}; \Xi^{(j)}) = \frac{f_H(y^{(j)}; \Xi^{(j)})}{F_H(\theta^{(j)}; \Xi^{(j)})}$ is the adequate right-truncated density of the head part of the marginal composite H -Inverse Weibull distribution truncated at the threshold point $\theta^{(j)}$. The expression for the mixing weight $r_{H,IW}^{(j)}$ can be obtained using (2). The dependency among the two types of claims, say $(y_i^{(1)}, y_i^{(2)})$, can be studied using the Gumbel copula as follows:

$$C_\phi(F_1(y_i^{(1)}), F_2(y_i^{(2)})) = \exp\left\{-\left[\left(-\log F_1(y_i^{(1)})\right)^\phi + \left(-\log F_2(y_i^{(2)})\right)^\phi\right]^{\frac{1}{\phi}}\right\}, \quad (7)$$

where $y_i^{(1)}$ and $y_i^{(2)}$ are the realizations for the two types of claims $Y^{(1)}$ and $Y^{(2)}$. $F_1(y_i^{(1)})$ and $F_2(y_i^{(2)})$ are the cdfs of composite H -Inverse Weibull models evaluated at $y_i^{(1)}$ and $y_i^{(2)}$, respectively. ϕ is the asymmetric parameter controlling the correlations among $(Y^{(1)}, Y^{(2)})$. We use three different parametric distributions for the H (head) part of the bivariate composite models, namely the Weibull distribution, the Paralogistic distribution, and the inverse Burr distribution. In this paper, for expository purposes, the development of the bivariate composite Weibull-Inverse Weibull (W-IW) distribution is presented.

Particular Case: The Bivariate Composite Weibull-Inverse Weibull (W-IW) Distribution

Let $Y_i^{(j)}$ be a random variable (r.v.) obtained by considering the Weibull distribution for the head and the Inverse Weibull distribution for the tail part of the marginal composite model. The pdf of the r.v. $Y_i^{(j)}$ can be written as

$$f_j(y_i^{(j)}) = \begin{cases} r_{W,IW}^{(j)} \frac{\frac{\mu^{(j)}}{\sigma^{(j)}} \exp\left\{-\left(\frac{y_i^{(j)}}{\sigma^{(j)}}\right)^{\mu^{(j)}}\right\} \left(\frac{y_i^{(j)}}{\sigma^{(j)}}\right)^{\mu^{(j)}-1}}{1 - \exp\left\{-\left(\frac{\theta^{(j)}}{\sigma^{(j)}}\right)^{\mu^{(j)}}\right\}} & \text{for } 0 < y_i^{(j)} \leq \theta^{(j)}, \\ (1 - r_{H,IW}^{(j)}) \frac{\frac{\alpha^{(j)}}{y_i^{(j)}} \left(\frac{\gamma^{(j)}}{y_i^{(j)}}\right)^{\alpha^{(j)}} \exp\left\{-\left(\frac{\gamma^{(j)}}{y_i^{(j)}}\right)^{\alpha^{(j)}}\right\}}{1 - \exp\left\{-\left(\frac{\gamma^{(j)}}{\theta^{(j)}}\right)^{\alpha^{(j)}}\right\}} & \text{for } \theta^{(j)} < y_i^{(j)} < \infty. \end{cases} \quad (8)$$

For $i = 1, 2, \dots, n$ and $j = 1, 2$, where $\mu^{(j)} > 0$, $\sigma^{(j)} > 0$, the scale parameter $\gamma^{(j)} > 0$, $\alpha^{(j)} > 0$, threshold point $\theta^{(j)} > 0$, and $r_{W,IW}^{(j)} \in [0, 1]$ is the mixing weight of the composite model. The analytical expression for the mixing weight $r_{W,IW}^{(j)}$ can be easily obtained using (2).

The cdf of the composite W-IW distribution is

$$F_j(y_i^{(j)}) = \begin{cases} r_{W,IW}^{(j)} \frac{1 - \exp\left\{-\left(\frac{y_i^{(j)}}{\sigma^{(j)}}\right)^{\mu^{(j)}}\right\}}{1 - \exp\left\{-\left(\frac{\theta^{(j)}}{\sigma^{(j)}}\right)^{\mu^{(j)}}\right\}} & \text{for } 0 < y_i^{(j)} \leq \theta^{(j)}, \\ r_{W,IW}^{(j)} + (1 - r_{W,IW}^{(j)}) \frac{\exp\left\{-\left(\frac{\gamma^{(j)}}{y_i^{(j)}}\right)^{\alpha^{(j)}}\right\} - \exp\left\{-\left(\frac{\gamma^{(j)}}{\theta^{(j)}}\right)^{\alpha^{(j)}}\right\}}{1 - \exp\left\{-\left(\frac{\gamma^{(j)}}{\theta^{(j)}}\right)^{\alpha^{(j)}}\right\}} & \text{for } \theta^{(j)} < y_i^{(j)} < \infty. \end{cases} \quad (9)$$

The joint cdf of the bivariate composite W-IW distribution can be obtained by coupling two marginal W-IW cdfs using the Gumbel copula, as shown in (4). Similarly, the joint cdf of the bivariate composite Paralogistic–Inverse Weibull (P-IW) distribution and the bivariate composite Inverse Burr–Inverse Weibull (IB-IW) distribution can also be obtained by coupling two marginal P-IW and IB-IW cdfs using the Gumbel copula given in (4).

4. Parameter Estimation via the IFM Method

The objective of this section is to explain how to use the maximum likelihood (ML) method to estimate the parameters related to both the marginals and the copula. Ref. [15] discussed a method known as the IFM for parameter estimation of the copula density, which is dependent on the knowledge of the marginals. This method involves two steps; first, the parameters of the marginal cdfs are estimated, and then the copula parameters are obtained by maximizing the likelihood function of the copula, with the marginal parameters replaced by estimators obtained in the first step. The steps involved in the IFM method for the estimation of the marginal model parameters as well as the copula parameter are given below.

- Step 1: Let $Y_1^{(j)}, Y_2^{(j)}, \dots, Y_n^{(j)}$ for $j = 1, 2$ be a random sample of two types of claims from the marginal composite H -Inverse Weibull distribution described in (5). Here, Θ is the parameter vector for the two marginal composite H -Inverse Weibull models. We utilize the ML estimation procedure to estimate the marginal distribution parameters. The goal of the ML estimation procedure is to find the parameter values that maximize:

$$l(\Theta^{(j)} | y_i^{(j)}) = \sum_{i=1}^n \ln(f_j(y_i^{(j)} | \Theta))$$

$$\begin{aligned}
l(\Theta^{(j)}|y_i^{(j)}) &= \sum_{i=1}^n \ln \left[r_{H,IW}^{(j)} \frac{f_H(y_i^{(j)}; \Xi^{(j)})}{F_H(\theta; \Xi^{(j)})} \mathbb{I}[y_i < \theta^{(j)}] \right. \\
&\quad \left. + (1 - r_{H,IW}^{(j)}) \frac{\frac{\alpha^{(j)}}{y_i^{(j)}} \left(\frac{\gamma^{(j)}}{y_i^{(j)}} \right)^{\alpha^{(j)}} \exp \left\{ - \left(\frac{\gamma^{(j)}}{y_i^{(j)}} \right)^{\alpha^{(j)}} \right\}}{1 - \exp \left\{ - \left(\frac{\gamma^{(j)}}{\theta^{(j)}} \right)^{\alpha^{(j)}} \right\}} \mathbb{I}[y_i^{(j)} \geq \theta^{(j)}] \right] \\
l(\Theta^{(j)}|y_i^{(j)}) &= \sum_{i=1}^n \left[\ln r_{H,IW}^{(j)} \mathbb{I}[y_i^{(j)} < \theta^{(j)}] + \ln f_H(y_i^{(j)}; \Xi^{(j)}) \mathbb{I}[y_i^{(j)} < \theta^{(j)}] \right. \\
&\quad - \ln F_H(\theta^{(j)}; \Xi^{(j)}) \mathbb{I}[y_i^{(j)} < \theta^{(j)}] + \ln(1 - r_{H,IW}^{(j)}) \mathbb{I}[y_i^{(j)} \geq \theta^{(j)}] \\
&\quad + \ln \alpha^{(j)} - \ln y_i^{(j)} + \alpha^{(j)} (\ln \gamma^{(j)} - \ln y_i^{(j)}) - \left(\frac{\gamma^{(j)}}{y_i^{(j)}} \right)^{\alpha^{(j)}} \mathbb{I}[y_i^{(j)} \geq \theta^{(j)}] \\
&\quad \left. - \ln \left(1 - \exp \left\{ - \left(\frac{\gamma^{(j)}}{\theta^{(j)}} \right)^{\alpha^{(j)}} \right\} \right) \mathbb{I}[y_i^{(j)} \geq \theta^{(j)}] \right] \quad (10)
\end{aligned}$$

For $j = 1, 2$, the non-overlapping density parts between the body and tail of a composite distribution make parameter estimation typically straightforward, allowing one to factor out the likelihood function and independently estimate the three components of the distribution—mixing weight, body, and tail. All parameters of the two marginal composite H -Inverse Weibull distributions are estimated using the numerical optimization tool `optim()`, included in the `stats` package of the R programming language.

- Step 2: The estimated parameters of the marginal df's are used to estimate the Gumbel copula parameter and to compute the value of the likelihood function associated with the dependence structure. To ease the process of finding the estimate of the copula parameter, we use the `rvinecopula` package of the R software 4.1.2 by passing the argument `gumbel` as the family.

5. Simulation Study

In this section, we present a simulation study for the proposed bivariate composite W-IW distribution using the Gumbel copula. Initially, we detail the generation of random samples from the bivariate composite W-IW distribution. The process involves employing the conditional procedure for random sample generation, as outlined in [16]. We consider Y_1, Y_2 be the random sample from bivariate composite W-IW distribution derived using Gumbel copula $C(u_1, u_2)$ (see (4)). The conditional distribution of the (U, V) is $P(V \leq v | U = u) = \frac{\partial}{\partial u} C(u_1, u_2)$. Utilizing this conditional distribution method, we generate the random numbers (y_1, y_2) from the proposed bivariate composite W-IW distribution. The detailed algorithm to generate the random numbers from the bivariate composite W-IW distribution is as follows:

Step	Random Numbers from Bivariate Composite W-IW Distribution
1.	Generate two independent samples, say u_1 and s from $U(0, 1)$.
2.	Set $s = \frac{\partial}{\partial u} C(u_1, u_2)$ and solve for u_2 .
3.	Find $y_1 = F_1^{-1}(u_1; \Theta^{(1)})$ and $y_2 = F_2^{-1}(u_2; \Theta^{(2)})$ where F_1^{-1} and F_2^{-1} are the quantile functions of the composite W-IW distribution having parameter space $\Theta^{(1)}$ and $\Theta^{(2)}$ at u_1 and u_2 , respectively.
4.	Finally, the required random sample is (y_1, y_2) .

Based on the following data obtained from the bivariate composite W-IW distribution, a simulation study was conducted and replicated 1000 times. We employed the IFM approach to estimate the parameters. The values of the parameters $\Theta^{(1)} = (\mu^{(1)}, \sigma^{(1)}, \alpha^{(1)}, \gamma^{(1)}, \theta^{(1)})$, $\Theta^{(2)} = (\mu^{(2)}, \sigma^{(2)}, \alpha^{(2)}, \gamma^{(2)}, \theta^{(2)})$ and Gumbel copula parameter ϕ are chosen for different

sample sizes $n = 50, 100, 200, 500$. In particular, we choose $\Theta^{(1)} = (2.5, 0.8, 3.8, 0.8, 2.3)$, $\Theta^{(2)} = (0.6, 0.5, 2.5, 0.5, 15)$ and $\phi = 1.5$. Figures 1 and 2 represent the box plots of bias and estimates of bivariate composite W-IW distribution for the sample sizes $n = 50, 100, 200, 500$. As the sample size increases, the estimators approach the true values, with smaller bias. Table 1 contains the Average Absolute Bias (AAB) and Root Mean Square Error (RMSE) of the parameter estimates of the bivariate composite W-IW distribution for the sample sizes $n = 50, 100, 200, 500$. It can be observed that the simulated AAB and RMSE decreases as sample size increases.

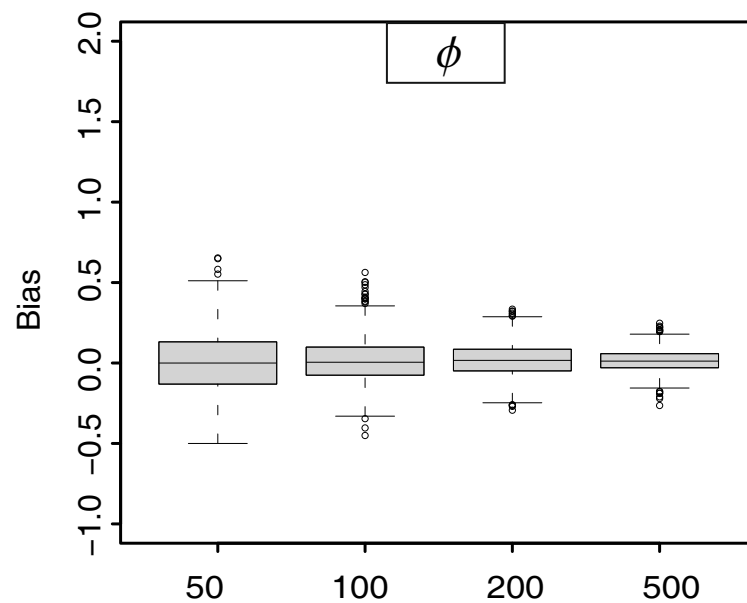


Figure 1. Box plot of bias of the parameter estimates of bivariate composite W-IW distribution with $\phi = 1.5$ obtained for different sample sizes $n = 50, 100, 200, 500$.

Table 1. Average Absolute Bias (AAB) and Root Mean Square Error (RMSE) of parameter estimates for bivariate composite W-IW distribution.

Parameters	n							
	50		100		200		500	
	AAB	RMSE	AAB	RMSE	AAB	RMSE	AAB	RMSE
$\mu^{(1)}$	0.030	0.521	0.020	0.472	0.009	0.451	0.006	0.388
$\sigma^{(1)}$	0.201	2.890	0.085	0.086	0.067	0.079	0.034	0.058
$\alpha^{(1)}$	5.698	14.515	6.057	14.507	5.350	11.043	5.142	9.779
$\gamma^{(1)}$	0.450	0.744	0.399	0.756	0.337	0.765	0.106	0.758
$\theta^{(1)}$	0.888	1.023	0.806	0.982	0.720	0.957	0.513	0.931
$\mu^{(2)}$	0.081	0.079	0.072	0.054	0.048	0.037	0.039	0.023
$\sigma^{(2)}$	0.087	0.170	0.058	0.169	0.015	0.090	0.013	0.041
$\alpha^{(2)}$	5.853	6.189	5.380	5.945	4.988	5.657	4.171	5.029
$\gamma^{(2)}$	4.627	9.982	3.568	6.798	3.029	4.945	2.565	3.579
$\theta^{(2)}$	29.663	40.394	27.324	36.842	24.638	33.542	17.759	28.960
ϕ	0.069	0.197	0.068	0.138	0.053	0.102	0.051	0.068

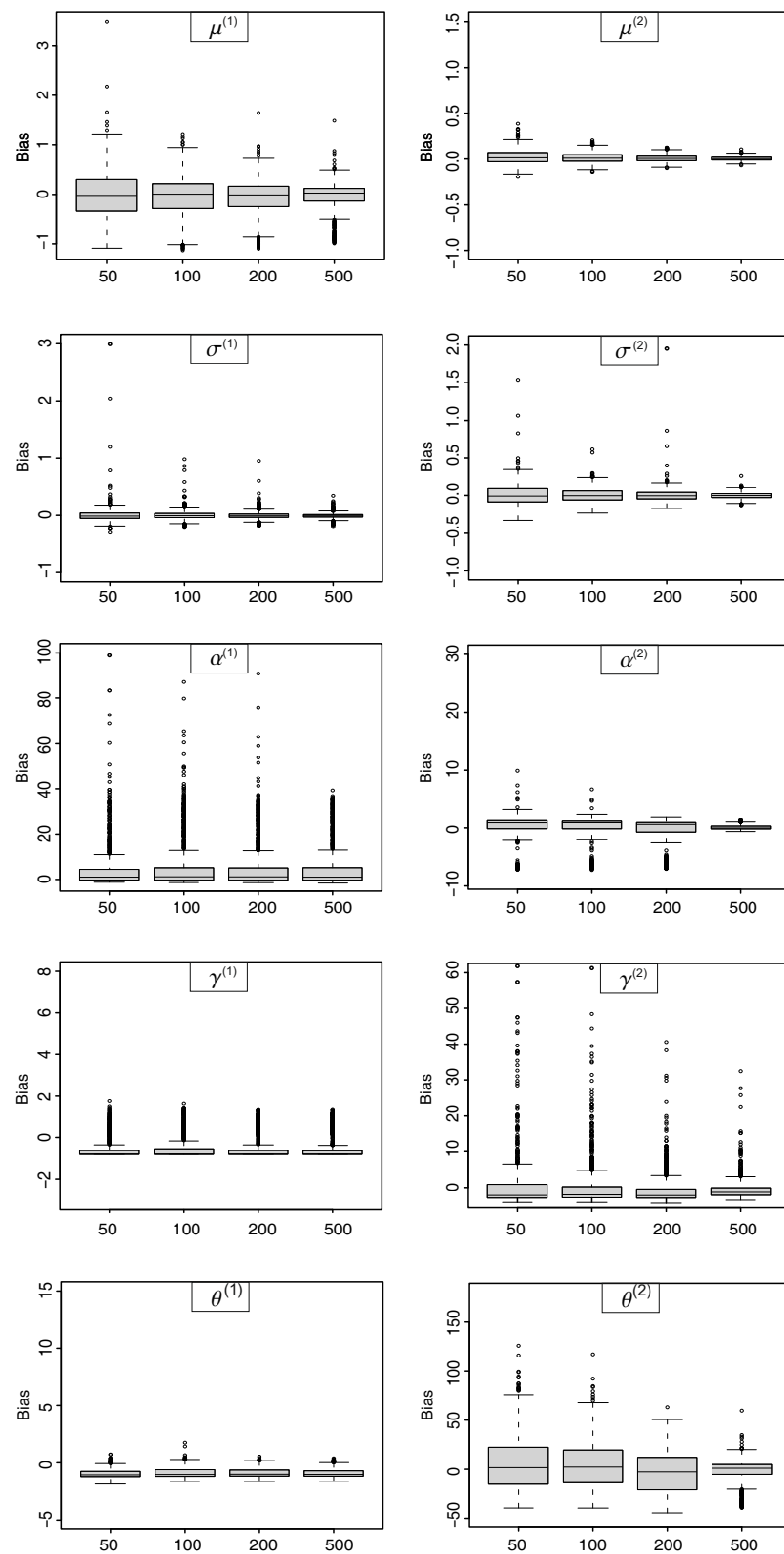


Figure 2. Box plot of bias of the parameter estimates of bivariate composite W-IW distribution with $\Theta^{(1)} = (2.5, 0.8, 3.8, 0.8, 2.3)$, $\Theta^{(2)} = (0.6, 0.5, 2.5, 0.5, 15)$ obtained for different sample sizes $n = 50, 100, 200, 500$.

6. Numerical Illustration

6.1. Dataset 1: Greece MTPL Dataset

In this section, we illustrate the proposed methodology using motor third-party liability (MTPL) insurance policies with non-zero property claims from the years 2012 to 2019. A major insurance firm in Greece generously provided the dataset for this study. The dataset contains 7263 motor vehicle insurance policies collected over the period from 2012 to 2019, all of which have complete records. The following section provides a detailed description of the MTPL dataset. The variables associated with the MTPL dataset are `tcost_bi` and `tcost_pd`, representing the cost of bodily injury claims and the cost of property damage claims, respectively. They are numeric vectors showing the total amount of bodily injury claims and property damage claims. Graphical representations of both types of claims, namely bodily injury claims and property damage claims, are presented in Figures 3–5. Both types of claims exhibit various peculiarities of insurance data, including positive skewness, unimodality, and tail heaviness. Figure 6, which displays the scatter plot of the MTPL dataset in natural logarithm scales, illustrates that the dependency is not linear. These data reveal extreme value dependence, i.e., the heavier the costs, the stronger the dependence. Additionally, it can be seen from Figure 6 that for small values of both claims, the dependence is almost zero, but the shape suggests positive dependence for larger costs of both claims. Clearly, this denotes a change in the joint distribution between the smaller and the larger claim costs. Here, both claims are known to be strongly correlated with high values but less correlated at low values; hence, the Gumbel copula will be an appropriate choice.

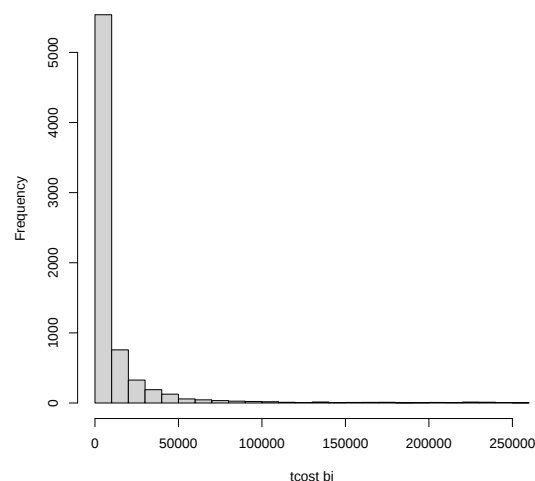


Figure 3. Histogram of bodily injury claims for the MTPL dataset.

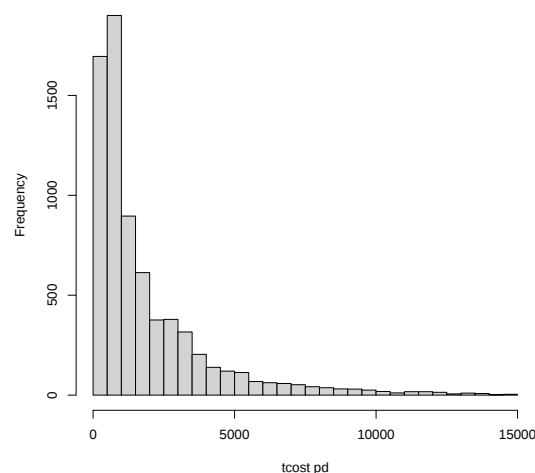


Figure 4. Histogram of property damage claims for the MTPL dataset.

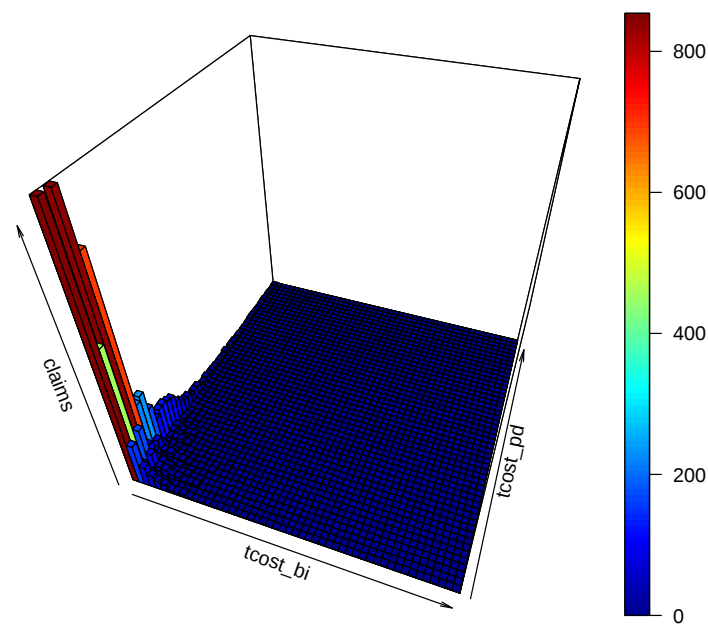


Figure 5. Histogram of bodily injury claims and property damage claims for the MTPL dataset.

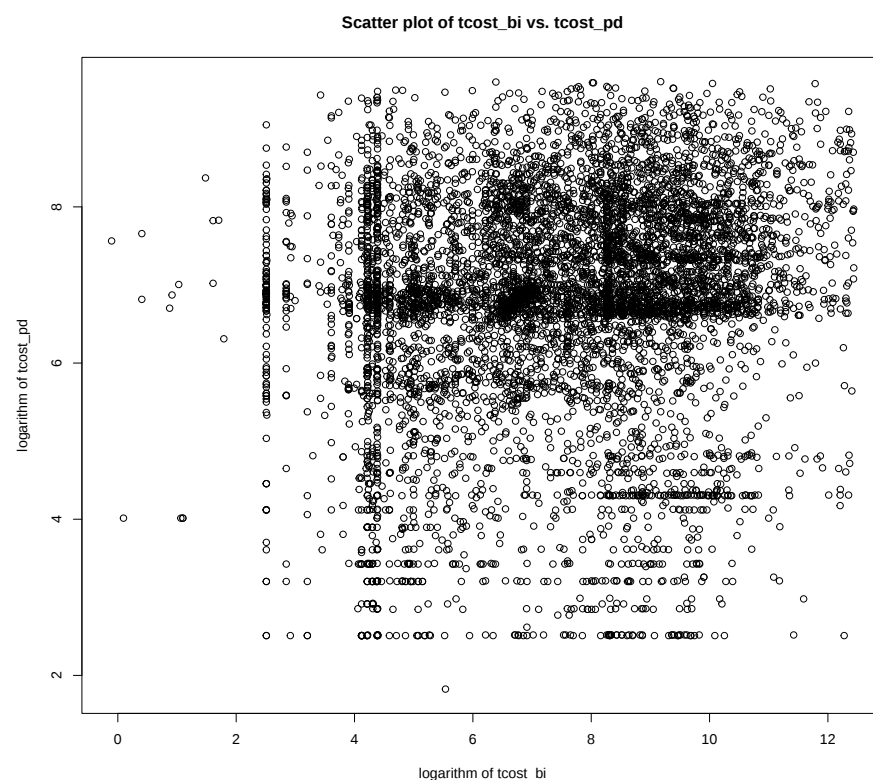


Figure 6. Scatter plot of bodily injury claims vs. property damage claims for the MTPL dataset.

6.2. Dataset 2: Danish Fire Insurance Dataset

The provided dataset describes fire insurance claims in Denmark, gathered from the Copenhagen Reinsurance Company for the period between 1980 and 1990. The dataset is available on the following website: <https://search.r-project.org/CRAN/refmans/fitdistrplus/html/danish.html> (accessed on 18 January 2024). It consists of three main components: loss to buildings, loss to contents, and loss to profit. However, our specific focus in this case revolves around modeling the interdependence between the first two components. The dataset comprises a total of 1501 observations. Our analysis specifi-

cally concentrates on cases where both losses have non-null values. In Figures 7–9, graphical representations of these claim types are provided. Notably, both *loss to buildings* and *loss to contents* exhibit several notable features commonly observed in insurance data, including a positively skewed distribution, a single mode (unimodality), and heavy tails indicating the presence of extreme values. The visual representation in Figure 10, which showcases a scatter plot of the Danish fire loss dataset using natural logarithmic scales, provides valuable insights into the relationship between these claims. It becomes apparent that the dependency between these claim types is not linear. Instead, the data suggest a non-linear relationship, and they reveal the existence of extreme value dependence. In other words, as claim costs increase, the strength of their interdependence becomes more pronounced. Moreover, a closer examination of Figure 10 reveals a fascinating pattern; for smaller values of both *loss to buildings* and *loss to contents*, the interdependence is nearly negligible, almost approaching zero. However, as losses grow larger, the figure clearly indicates a positive dependence emerging. This shift in the joint distribution, from near-zero dependence for smaller losses to a positive relationship for larger losses, represents a significant and noteworthy change in the nature of the joint distribution between these two categories of losses.

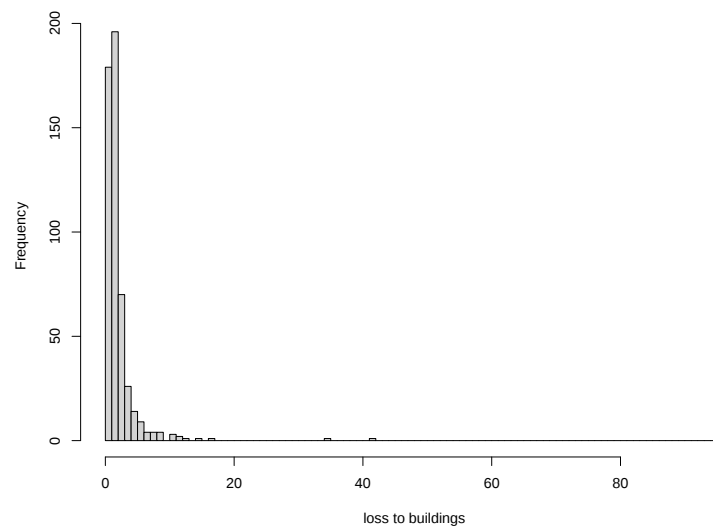


Figure 7. Histogram of *loss to buildings* for the Danish dataset.

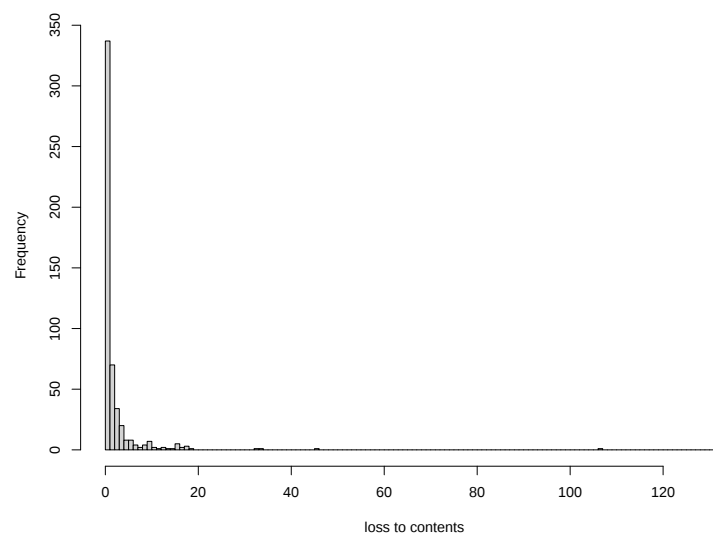


Figure 8. Histogram of *loss to contents* for the Danish dataset.

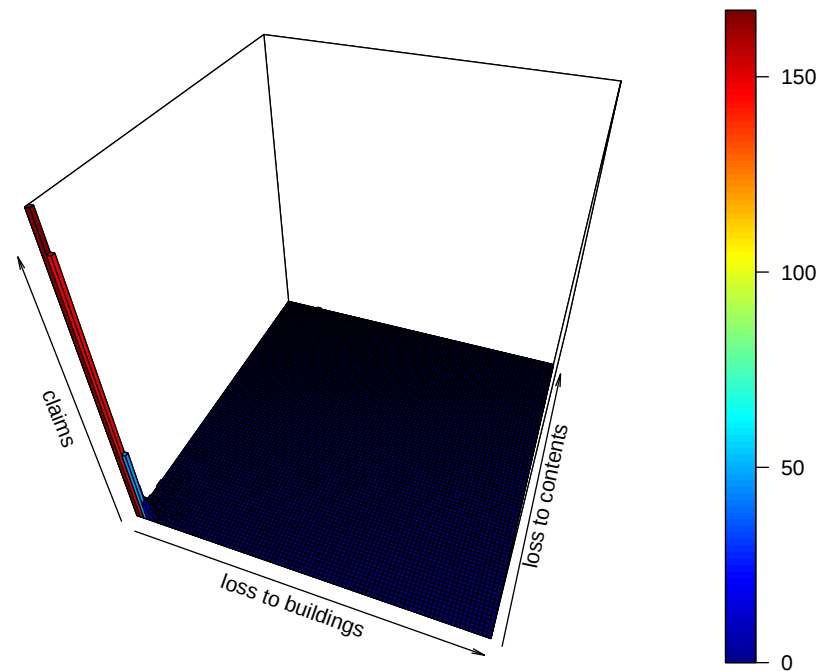


Figure 9. Histogram of loss to buildings and loss to contents for the Danish dataset.

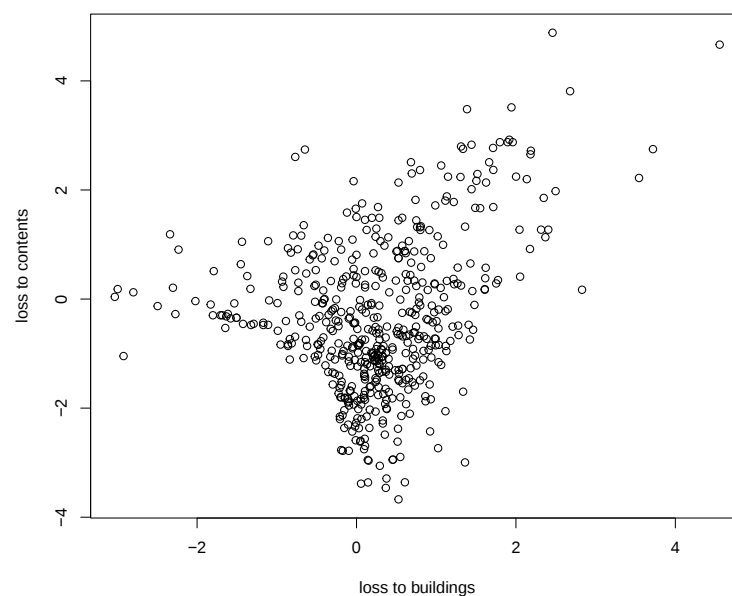


Figure 10. Scatter plot of loss to buildings vs. loss to contents for the Danish dataset.

6.3. Model Fitting

The fitting of the bivariate composite distributions, namely bivariate composite Weibull–Inverse Weibull (W-IW), the bivariate composite Paralogistic–Inverse Weibull (P-IW) distribution, and the bivariate composite Inverse Burr–Inverse Weibull (IB-IW) distribution, as well as other existing bivariate distributions, such as bivariate Lomax (BL), bivariate Mardia’s Pareto Type I (BMPI), and bivariate Burr (BB) model, are examined based on the MTP1 dataset in Section 6.1 and Danish fire loss dataset in Section 6.2. In this section, we present the results based on two model selection criteria for the proposed distribution. Tables 2 and 3 provides the values of negative log likelihood (NLL), Akaike’s information criterion (AIC), and Bayesian information criterion (BIC) for the MTP1 and Danish fire loss datasets, respectively. The formulas involved in the computation of the above-mentioned model selection criteria, namely AIC and BIC, are:

$$AIC = -2\mathcal{L}(\hat{\Theta}) + 2 \times \text{df},$$

where $\mathcal{L}(\hat{\Theta})$ is the maximum log-likelihood and $\hat{\Theta}$ is the vector of the estimated model parameters.

$$BIC = -2\mathcal{L}(\hat{\Theta}) + \log(n) \times \text{df},$$

where n is the sample size of the dataset and df is the number of fitted parameters of the distribution.

Note that for NLL, AIC , and BIC , smaller values indicate a better fit of the distribution to the empirical data. Tables 2 and 3 show that the bivariate composite W-IW distribution performs better than the remaining bivariate composite distributions and other existing bivariate distributions for the MTPL dataset and bivariate composite IB-IW distribution performs better than the remaining bivariate composite distributions and other existing bivariate distributions for the Danish fire loss dataset. Tables 4 and 5 present the parameter estimates (marginal model parameters and Gumbel copula parameter) of the bivariate composite distributions as well as existing bivariate distributions for the MTPL dataset and Danish fire loss dataset, respectively.

Table 2. Values of NLL, AIC , and BIC for the MTPL dataset for competing distributions.

Model	Parameters	NLL	AIC	BIC
W-IW	9	132,460.65	264,939.31	265,001.32
P-IW	9	132,673.95	265,365.91	265,428.01
IB-IW	11	133,418.21	266,858.41	266,934.21
BL	3	134,030.91	268,067.81	268,088.47
BMPI	3	143,634.61	287,275.21	287,295.88
BB	3	146,181.11	292,368.21	292,388.88

Table 3. Values of NLL, AIC , and BIC for the Danish dataset for competing distributions.

Model	Parameters	NLL	AIC	BIC
W-IW	9	3731.75	7481.51	7531.71
P-IW	9	3738.05	7494.11	7541.94
IB-IW	11	3719.91	7461.82	7524.31
BL	3	4119.54	8245.08	8261.02
BMPI	3	4018.51	8043.02	8058.96
BB	3	4283.83	8573.66	8589.61

Table 4. Parameter estimates of competing distributions for the MTPL dataset.

Parameter	W-IW	P-IW	IB-IW	BL	BMPI	BB
$\mu^{(1)}$	0.5394	0.7953	6.8642	-	-	-
$\sigma^{(1)}$	-	-	0.3771	-	-	-
$\alpha^{(1)}$	1.3949	0.8948	1.7376	-	-	-
$\gamma^{(1)}$	13,518.1554	10,034.7245	27,070.5612	-	-	-
$\theta^{(1)}$	23,926.6624	25,449.2312	29,407.7415	-	-	-
$\mu^{(2)}$	0.8751	1.2554	0.5485	-	-	-
$\sigma^{(2)}$	-	-	1.2214	-	-	-
$\alpha^{(2)}$	13.5462	2.4359	15.5139	-	-	-
$\gamma^{(2)}$	25,843.0438	11,634.1078	11,410.9912	-	-	-
$\theta^{(2)}$	43,134.3912	16,941.3751	13,100.0011	-	-	-
ϕ	1.1141	1.0711	1.0161	-	-	-
a_1	-	-	-	1.5467	0.1279	487.8339
a_2	-	-	-	0.0001	1.1112	7772.1023
p	-	-	-	0.0005	1.7805	1185.0597

Table 5. Parameter estimates of competing distributions for the Danish fire loss dataset.

Parameter	W-IW	P-IW	IB-IW	BL	BMPI	BB
$\mu^{(1)}$	2.5648	2.6956	0.0641	-	-	-
$\sigma^{(1)}$	-	-	34.1352	-	-	-
$\alpha^{(1)}$	1.7463	1.7778	1.7931	-	-	-
$\gamma^{(1)}$	0.5164	0.6061	0.7825	-	-	-
$\theta^{(1)}$	1.4341	1.5201	1.0581	-	-	-
$\mu^{(2)}$	1.7291	1.8079	0.8502	-	-	-
$\sigma^{(2)}$	-	-	1.8846	-	-	-
$\alpha^{(2)}$	1.0369	1.0471	1.1146	-	-	-
$\gamma^{(2)}$	0.3018	0.3176	0.4047	-	-	-
$\theta^{(2)}$	0.4323	0.4765	0.1996	-	-	-
ϕ	1.1800	1.1766	1.1767	-	-	-
a_1	-	-	-	3.0860	0.2579	0.9649
a_2	-	-	-	0.2274	0.9874	1.3341
p	-	-	-	0.4048	1.9874	1.3687

6.4. Analysis of Dependence

To examine the goodness of fit of the fitted distribution in terms of dependence modelling, we provide the values of Kendall's tau (τ), the upper-tail dependence parameter λ_U for the bivariate composite distributions generated using the Gumbel copula. In the empirical examination, the calculated τ stands at 0.099 for the MTPL dataset and 0.085 for the Danish fire loss dataset, indicating the observed ordinal correlation in each respective dataset. In case of the MTPL dataset, Table 6 indicates that the Kendall's tau for the W-IW bivariate composite distributions matches the empirical τ reasonably well, implying that the suggested Gumbel copula model can properly reflect the dependence structure of both types of claims in the MTPL dataset. In case of Danish fire loss dataset, the fitted values of the Kendall's tau for the proposed bivariate composite distributions are far away from the empirical value of the Kendall's tau.

Table 6 also presents the values of the upper-tail dependence parameter (λ_U) for both the datasets. The empirical values of the upper-tail dependence for the MTPL dataset and Danish fire loss dataset are 0.133 and 0.198, respectively. From Table 6, it is observed that for the proposed distributions, upper-tail dependence parameter $\lambda_U \in (0, 1]$, indicating that bivariate composite distributions exhibit upper-tail dependence for both the datasets. Table 6 shows the empirical upper-tail dependence for the MTPL dataset coincides with the upper-tail dependence of the proposed bivariate composite W-IW distribution, indicating the adequate fit of the bivariate composite W-IW distribution to the extremely large values of the MTPL dataset. In case of Danish fire loss dataset, the fitted value of Kendall's tau based on proposed bivariate composite distributions lies far away from its empirical counterpart, but the value of fitted upper-tail dependence matches the empirical value of upper-tail dependence. It suggests a situation where the proposed bivariate distributions captures certain aspects of dependence (specifically, tail dependence) but fails to accurately represent the overall Kendall's tau.

Table 6. Dependence measures of the bivariate composite distributions for the MTPL dataset and Danish fire loss data set dataset.

Model	MTPL		Danish	
	τ	λ_U	τ	λ_U
W-IW	0.102	0.137	0.153	0.199
P-IW	0.066	0.089	0.150	0.197
IB-IW	0.015	0.021	0.151	0.198

7. Conclusions

In this paper, we presented a novel family of bivariate composite distributions for simultaneously modelling small and/or moderate and large costs from different types of claims. The detailed methodology involved in the development of bivariate composite distribution is presented. A Gumbel copula is utilized to specify the correlation structure between two types of random variables. For expository purposes, the genesis of bivariate composite Weibull–Inverse Weibull (W-IW) distribution via Gumbel copula is exhibited. Application of the proposed new class of distribution is illustrated using the MTPL bodily injury and property damage dataset and Danish fire loss dataset. The numerical results obtained show reasonably good results for the bivariate composite distributions as compared to other existing bivariate distributions. This suggests that the proposed distributions offers a competitive and potentially more accurate representation of the joint distribution of insurance claim costs. A possible limitation of the proposed study is the choice of the copula. The choice of the Gumbel copula may impact the results, and its appropriateness depends on the underlying dependence structure. Sensitivity analyses or comparisons with alternative copulas could provide insights into the robustness of the proposed distributions. A potentially fruitful line of further research is to consider the influence of individual and coverage-specific covariates on the parameter(s) of the proposed class of distribution by introducing regression components into the composite marginal models.

Author Contributions: Conceptualization, D.B., G.A. and G.T.; methodology, D.B. and G.A.; software, D.B. and G.A.; validation, D.B., G.A. and G.T.; formal analysis, D.B. and G.A.; investigation, D.B. and G.A.; resources, D.B., G.A. and G.T.; data curation, D.B., G.A. and G.T.; writing—original draft preparation, G.A.; writing—review and editing, D.B., G.A. and G.T.; visualization, G.A.; supervision, D.B. and G.T.; project administration, D.B.; funding acquisition, G.T. All authors have read and agreed to the published version of the manuscript.

Funding: The present work is a part of project granted by Department of Science and Technology, Government of India under the Core Research Grant scheme (CRG/2019/002993). D.B. and G.A. thank the funding agency for financial support.

Data Availability Statement: Restrictions apply to the availability of Dataset 1. Dataset 1 was obtained from third party and is available with the permission of the third party. Dataset 2 is a publicly available dataset. These data can be found here: [<https://search.r-project.org/CRAN/refmans/fitdistrplus/html/danish.html>] (accessed on 18 January 2024)].

Acknowledgments: We acknowledge Ravindra Khattree, Oakland University, Rochester, USA for the initial discussions and careful reading of the first draft of the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Bakar, S.A.A.; Hamzaha, N.A.; Maghsoudia, M.; Nadarajah, S. Modelling loss data using composite models. *Insur. Math. Econ.* **2015**, *61*, 146–154. [[CrossRef](#)]
2. Cooray, K.; Ananda, M.M. Modelling actuarial data with a composite lognormal-Pareto model. *Scand. Actuar. J.* **2005**, *5*, 321–334. [[CrossRef](#)]
3. Scollnik, D. On composite lognormal-Pareto models. *Scand. Actuar. J.* **2007**, 20–33. [[CrossRef](#)]
4. Ciumara, R. An actuarial model based on the composite Weibull-Pareto distribution. *Math. Rep.* **2006**, *8*, 401–414.
5. Scollnik, D.P.M.; Sun, C. Modeling with Weibull-Pareto models. *N. Am. Actuar. J.* **2012**, *16*, 260–272. [[CrossRef](#)]
6. Nadarajah, S.; Bakar, S. New composite models for the Danish fire insurance data. *Scand. Actuar. J.* **2014**, *2014*, 180–187. [[CrossRef](#)]
7. Grün, B.; Miljkovic, T. Extending composite loss models using a general framework of advanced computational tools. *Scand. Actuar. J.* **2019**, *8*, 642–660. [[CrossRef](#)]
8. Wang, Y.; Haff, I.H.; Huseby, A. Modelling extreme claims via composite models and threshold selection methods. *Insur. Math. Econ.* **2020**, *91*, 257–268. [[CrossRef](#)]
9. Benatmane, C.; Zeghdoudi, H.; Shanker, R.; Lazri, N. Composite Rayleigh-Pareto distribution: Application to real fire insurance losses data set. *J. Stat. Manag. Syst.* **2020**, *24*, 545–557. [[CrossRef](#)]
10. Benchettah, M.H.; Zeghdoudi, H.; Raman, V. On composite length-biased exponential-Pareto distribution: Properties, simulation, and application in actuarial science. *Front. Appl. Math. Stat.* **2023**, *9*, 545–557. [[CrossRef](#)]

11. Fung, T.C.; Tzougas, G.; Wüthrich, M.V. Mixture composite regression models with multi-type feature selection. *N. Am. Actuar. J.* **2022**, *27*, 396–428. [[CrossRef](#)]
12. Reynkens, T.; Verbelen, R.; Beirlant, J.; Antonio, K. Modelling censored losses using splicing: A global fit strategy with mixed Erlang and extreme value distributions. *Insur. Math. Econ.* **2017**, *77*, 65–77. [[CrossRef](#)]
13. Frees, E.W.; Valdez, E.A. Understanding relationships using copulas. *N. Am. Actuar. J.* **1998**, *2*, 1–25. [[CrossRef](#)]
14. Cherubini, U.; Luciano, E.; Vecchiato, W. *Copula Methods in Finance*; John Wiley & Sons: Chichester, UK, 2004.
15. Joe, H. *Multivariate Models and Dependence Concepts*, 1st ed.; CRC Press: New York, NY, USA, 1997.
16. Nelsen, R.B. *An Introduction to Copulas*; Springer: Berlin/Heidelberg, Germany, 2006.
17. Nikoloulopoulos, A.K.; Joe, H.; Li, H. Vine copulas with asymmetric tail dependence and applications to financial return data. *Comput. Stat. Data Anal.* **2012**, *56*, 3659–3673. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.