Copula Modeling: An Application to Workers' Compensation Claims

A Major Qualifying Project Report submitted to the faculty of WORCESTER POLYTECHNIC INSTITUTE in partial fulfillment of the requirements for the Degree of Bachelor of Science.

BY:
Lexi Ferrini
Alison Lambert
Donovan Robillard

DATE:
April 26, 2022

SUBMITTED TO:

Jon Abraham and Barry Posterro
Worcester Polytechnic Institute

Steve Bunker, Voon Lai
The Hanover Insurance Group

Abstract

Copulas are multivariate probability distributions used in the modeling of multiple random variables. In insurance, they are used to create models that preserve the relationship between a claim's loss amount and any associated expenses, especially in large loss scenarios. The goal of this project was to develop a copula model for losses and their associated expenses and determine whether their use produces different results than current modeling methods. Through data analysis and simulation, the team identified that a copula model could be applied to claims in Workers' Compensation. It was found that the copula model did not produce significantly different results than those produced using the sponsor's traditional methods, validating their current models.

Executive Summary

A copula is a multivariate function made up of any number of random variables where each marginal follows a Uniform (0,1) distribution. Any multivariate CDF can be written as a copula, which makes them an extremely useful mathematical tool when modeling multiple random variables, especially for variables that exhibit dependence (Klugman et al). Copulas prove extremely useful in the insurance industry. This project, sponsored by The Hanover Insurance Group, focused on introducing copulas into their modeling process for large losses, and then comparing the results with those from their traditional model. Currently, The Hanover separates large losses from smaller, or attritional, losses and then models the total losses and expenses from each of these claims as one single severity distribution. The Hanover wanted to see if modeling large losses and their associated expenses as individual distributions aggregated with a copula provided significantly different results. The Hanover provided the team with loss and premium data from three commercial lines of business. The team then created a copula model using Workers' Compensation data.

Using a variety of dependency measures, statistical tests, and visual indicators, the team determined that Workers' Compensation losses and expenses have a dependency structure best represented by the Heavy Right Tail (HRT) copula. The team was able to find the appropriate marginal distributions for losses and expenses, simulate values using the HRT copula, and provide statistical estimates for the top ten percent and top five percent of total losses and expenses. They found that the copula model does not offer statistically different results than the traditional model for large loss claims. We conclude that The Hanover's current modeling methods are appropriate for this data. The team is hopeful that with future improvements, this process could be applied to other business lines and yield additional interesting results.

Table of Contents

List of Tables

List of Figures

Introduction

Univariate distributions are a familiar topic to anyone who has studied probability or statistics at even the most basic level. There are examples of these distributions all around us, such as determining the average grade on a test or predicting how long it will take for a train to arrive. Many people also have experience with bivariate distributions, for example, modeling a person's height with their age, or identifying the correlation between a person's height and weight. Even those who have not studied mathematics are familiar with the idea of correlation between two variables: taller people generally weigh more (positive correlation), and students who miss more school days usually see lower grades (negative correlation). It is also possible to have two variables that have no relationship. An example of this could be when flipping a fair coin. An observer may obtain heads on the first toss, but on any subsequent attempt, the observer will have the same probability of heads, implying that having information about earlier tosses will not impact the prediction of the current toss. However, in situations where the data are dependent on each other in a consistent manner throughout their entire distributions, one can make assumptions about how two variables will move and work together.

These basic concepts of correlation and dependence are all extremely useful in modeling and predicting outcomes in the natural world. But what if one were to encounter situations where these simple models are not enough? Sometimes, the data contains multiple variables (more than two) that are dependent on one another, or maybe the way these variables are correlated is not consistent throughout. Consider one of the most common tasks faced by investors: stock portfolio diversification. To mitigate losses, they want to choose stocks that are uncorrelated, or those that have little chance of declining at similar times. But, in an extreme economic downturn, is this strategy enough to ensure that all stocks will not fall at the same time? Or consider the example of an insurance company providing Worker's Compensation and Homeowner's insurance. These two lines of business are not typically associated with each other. But, if there is a part of the country with a significant amount of business that experiences some sort of unexpected natural disaster, the insurance company will see large losses in both business lines. Copulas can be useful in these situations.

Background

Measures of Dependence

Suppose it is known that two random variables under investigation exhibit a relationship to one another, but it is not known exactly how they are related. To identify the relationship between these variables, one needs to understand how to calculate various measures of dependence. Measures of dependence can be used to quantify different characteristics of the relationship between two variables, such as linearity, monotonicity, and the occurrence of extreme values in the tails (Klugman et al). Here, four measures of dependence, Pearson's correlation coefficient, Spearman's Rho, Kendall's Tau, and tail dependence, will be discussed.

Pearson's Correlation Coefficient

The most common dependency measure is Pearson's correlation coefficient (r). Pearson's correlation coefficient takes values between -1 and 1 and measures the linear correlation between two data sets (Klugman et al). The Pearson correlation coefficient for random variables X_1 and X_2 is defined as the following:

$$r(X_1, X_2) = \frac{E(X_1X_2) - E(X_1)E(X_2)}{(Var(X_1)Var(X_2))^{1/2}}$$

Given empirical data, Pearson's correlation coefficient can be calculated as:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where x_i is the ith value of vector **x**, $\bar{x}$ is the mean of vector **x**, y_i is the ith value of vector **y**, and $\bar{y}$ is the mean of vector **y**.

As an example, consider the scenario where a teacher records the number of hours their students spend studying for a test, and compares it with how they perform on the test. Table 1 displays each students' total study hours, as well as their test grade. For this data, r = 0.962, is very close to 1. This indicates a strong, positive, linear relationship, which means that students who spent more hours studying did better on the test.

This idea can be extended to more than two variables, where instead of having a single value for r, there would be a matrix of different correlations between all the combination of variables. Some multivariable distributions such as the multivariate normal and **t**-distributions utilize a correlation matrix, along with other information about these random variables, to create joint densities. In many cases, however, linear correlation may not be a sufficient measure of dependence (Klugman et al). An example of this is when two variables exhibit weak relationships at some areas in the distribution, and very strong relationships in other sections. In

these cases, Pearson's correlation coefficient may give a value near 0, implying that there is no relationship between the two variables. Nevertheless, these variables may still depend on each other at various points throughout the distribution, and alternative measures of dependence need to be investigated.

Student ID	Hours Spent Studying	Test Grade
1	3.75	94
2	0.75	77
3	1.75	80
4	2.25	82
5	3.25	95
6	1.00	71
7	2.00	79
8	2.50	85
9	0.50	70
10	3.50	90
11	1.25	74
12	2.75	91
13	0.25	68
14	4.00	96
15	1.50	81
16	0.00	60
17	3.00	59

Table 1: Hours Spent Studying and Test Grades, $r = 0.962$

Spearman's Rho

When considering more complex data sets that are not independent, but also not evenly, linearly correlated throughout, new measures of dependency must be introduced. The first of those is Spearman's Rho (ρ_S). Spearman's Rho measures the strength and direction of the monotonic relationship between two ranked variables. A monotonic relationship between two variables means that as one variable moves up or down, the other will move in the same direction, and vice versa. Pearson's correlation coefficient also accounts for monotonicity, inherent in linear relationship between variables. However, Spearman's Rho is only concerned with monotonicity, so the underlying form of the monotonicity is unimportant. This means that two variables could exhibit a perfect linear relationship, an exponential relationship, or any other monotonic relationship, and still have the same values of Spearman's Rho. Pearson's correlation coefficient would be very different in these cases. Spearman's Rho relies on the rank (or percentile) of the observations from each variable, which allows this value to be interpreted as the correlation of the percentiles of the random variables.

Definition: Consider two random variables X_1 and X_2 with CDFs $F_1(X_1)$ and $F_2(X_2)$, then Spearman's Rho is defined as the following.

$$\rho_S = r(F_1(X_1), F_2(X_2))$$

Let U and V be two Uniform (0,1) random variables, then the statement above is equivalent to

$$\rho_S = r(U, V)$$

In practice, with empirical paired data, Spearman's Rho can be calculated as:

$$\rho_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

where d_i = the difference in rank of the ith pair, n = the number of pairs

Now, return to the test grade example, where Spearman's Rho can be calculated from the same data. Table 2 displays each students' total study hours and test grade, as well as the rank of each. The rank of a data point is its placement within the data. So, the highest test grade will be ranked at number one, while the lowest test grade will be the lowest ranked. The Spearman's Rho for this data is, ρ_S = 0.966. This is similar to the Pearson's correlation coefficient, which is reasonable considering that Spearman's Rho and Pearson's correlation coefficient each account for monotonicity.

The simplest example of the difference between Pearson's correlation coefficient and Spearman's Rho is to consider the relationship between X and X^2 for all $X > 0$. There is no linear relationship between these two variables, but there is a perfect monotonic relationship.

Student ID	Hours Spent Studying	Test Grade	Hours Spent Studying Rank	Test Grade Rank	Difference in Rank
1	3.75	94	2	3	1
2	0.75	77	14	12	2
3	1.75	80	10	10	0
4	2.25	82	8	8	0
5	3.25	95	4	2	2
6	1.00	71	13	14	1
7	2.00	79	9	11	2
8	2.50	85	7	7	0
9	0.50	70	15	15	0
10	3.50	90	3	5	2
11	1.25	74	12	13	1
12	2.75	91	6	4	2
13	0.25	68	16	16	0
14	4.00	96	1	1	0
15	1.50	81	11	9	2
16	0.00	60	17	17	0
17	3.00	59	5	6	1

Table 2: Hours Spent Studying and Test Grades with Ranks, $\rho_S = 0.966$

Kendall's Tau

Another measure of dependency that is reliant on the monotonicity of the relationship between two variables is Kendall's Tau. Kendall's Tau is an important parameter in many copulas and is often more efficient to calculate than Spearman's Rho (Klugman et al). This measure is an alternative to Spearman's Rho, especially when one has a small sample size with many tied ranks. Kendall's Tau (τ_k) is defined as the following:

Consider two identically distributed bivariate random variables (X_1, X_2) and (Y_1, Y_2), where X_1 and Y_1 come from the same distribution $F_1(x)$, and X_2 and Y_2 come from the same distribution $F_2(x)$. Then,

$$\tau_k(X_1, X_2) = P[(X_1 - Y_1)(X_2 - Y_2) > 0] - P[(X_1 - Y_1)(X_2 - Y_2) < 0]$$

This is equivalent to finding the expected value of the sign of $(X_1 - Y_1)(X_2 - Y_2)$. Kendall's Tau can be calculated from empirical data using the principles of concordance and discordance. For a pair of data points to be concordant, their rank must move in the same direction. That is, if $(x_1 - y_1)(x_2 - y_2) > 0$, then the pair is concordant. On the other hand, if $(x_1 - y_1)(x_2 - y_2) < 0$, then the pair is discordant, so the ranks of the pair move in opposite directions. The formula for Kendall's Tau, using these definitions, is:

$$\tau_k = \frac{C - D}{C + D}$$

where C = the total number of concordant pairs, D = the total number of discordant pairs.

To estimate Kendall's Tau from a data set, consider the study hours and test grade example once again, but with new, non-linear data. Table 3 provides values for study hours and test grades, the ranks of each, and the number of concordant and discordant pairs. With this data, r = 0.132, which indicates non-linearity. However, τ_k = 0.867, indicating that study hours and grades are still correlated. This relationship would have been missed without Kendall's Tau. Kendall's Tau is typically used in these cases; when the relationship between two variables doesn't follow a linear correlation, the two variables of interest are continuous with outliers, or the variables are ordinal.

Student ID	Hours Spent Studying	Test Grade	Hours Spent Studying Rank	Test Grade Rank	Concordant Pairs	Discordant Pairs
1	3.75	96	2	1	16	0
2	0.75	94	14	3	14	1
3	1.75	90	10	5	12	2
4	2.25	95	8	2	12	0
5	3.25	89	4	6	11	1
6	1	91	13	4	11	0
7	2	85	9	7	10	0
8	2.5	82	7	8	9	0
9	0.5	79	15	11	6	2
10	3.5	80	3	10	6	1
11	1.25	81	12	9	6	0
12	2.75	74	6	13	4	1
13	0.25	71	16	14	3	1
14	4	77	1	12	3	0
15	1.5	70	11	15	2	0
16	0	68	17	16	1	0
17	3	60	5	17	0	0

Table 3: Hours Spent Studying and Test Grades with Ranks, Concordant Pairs and Discordant Pairs, τ_k = 0.867

Tail Dependence

A fourth measure of dependence that proves useful with non-linear data, and particularly when evaluating extreme values, is tail dependence. The upper tail dependence is a measure of how severe one random variable is as given another is very severe (Klugman et al). The formulas for upper and lower tail dependence are:

$$\lambda_{Upper} = \lim_{u \to 1} P[U > z \mid V > z]$$

$$\lambda_{Lower} = \lim_{u \to 0} P[U \leq z \mid V \leq z]$$

where U and V are Uniform (0,1) random variables.

This measure of dependence separates scales from the underlying distributions, and only considers how severe one distribution is compared to the severity of another distribution. To illustrate this, return to the stock market diversification example mentioned earlier. Typically, an investor tries to diversify their portfolios with two stocks that are uncorrelated. However, during "normal" times these stocks may seem uncorrelated, when there are large drops in the market, one may see the price of both stocks drop significantly, which would imply some sort of dependence between the two (Klugman, et al). In this situation, where the goal is to be prepared for large loss scenarios, tail dependence may be the most important measure.

Summary

As discussed earlier, it is always important to consider more than one dependence measure, because a stock trader still wants to create a portfolio that will perform well during normal times. For this reason, they may still want to consider Pearson's Correlation to find uncorrelated stocks. There are also many other scenarios where Kendall's Tau or Spearman's Rho are the best indicators of the dependency of the data (Klugman et al). All these measures of dependence offer important insight into the underlying data structure. Thus, evaluating and reporting these measures for a given data set is the first step in developing a copula model.

Definition of a Copula and Sklar's Theorem

Definition: Let $U_1, \cdots, U_d$ be Uniform (0,1) random variables. Then, copula C is defined to be

$$C(U_1, \cdots, U_d) = P(U_1 \leq u_1, \cdots, U_d \leq u_d).$$

Essentially, a copula is a multivariate probability distribution which is made up of correlated percentiles. This definition can be extended using Sklar's Theorem, which allows this copula to be written as a multivariate distribution.

Sklar's Theorem: Let F be a d-dimensional CDF of the random vector $X = X_1, \cdots, X_d$ and marginal distributions $F_1, \cdots, F_d$. Then, there exists copula C such that

$$F(x_1, \cdots, x_d) = C[F(x_1), \cdots, F(x_d)] = P(U_1 \leq u_1, \cdots, U_d \leq u_d)$$

where $U_1, \cdots, U_d$ are Uniform (0,1) random variables.

The result of this theorem is that any multivariate distribution can be written as a copula of univariate marginal distributions. This allows the modeler to create a multivariate distribution given a set of marginal distributions and some idea about the underlying data structure. Using copulas, the formulas for the four important dependency measures can be rewritten as follows.

$$\rho_S = 12 \int_0^1 \int_0^1 C(u, v) du dv \; - \; 3$$

$$\tau_K = \; 4 \int_0^1 \int_0^1 C(u, v) c(u, v) du dv \; - \; 1$$

where $c(u, v) = \frac{\partial^2 C(u,v)}{\partial u \partial v}$, the density function.

$$\lambda_{Upper} = \lim_{u \to 1} \frac{1 - 2u + C(u, u)}{1 - \; P(V \; \leq \; u)}$$

$$\lambda_{Lower} = \lim_{u \to 0} \frac{C(u, u)}{u}$$

The Inverse Transform Method

The inverse transform method, one of the crucial results needed in the proof of Sklar's theorem, is stated below.

Theorem: Consider a random variable X that has CDF $F_X(x)$. Let U be a new random variable that is defined by the transformation $U = F_X(X)$. Then, U ~ Uniform (0,1), and $P(U \leq u) = u$ for $0 < u < 1$.

Proof: $F_X(X)$ is a function from $X \rightarrow (0,1)$. It is known for realizations of U, u, that $0 < u < 1$.

$$P(U \leq u) = P(F_X(X) \leq u)$$
$$= P(F_X^{-1}[F_X(X)] \leq F_X^{-1}(u))$$
$$= P(X \leq F_X^{-1}(u))$$
$$= F_X\,[F_X^{-1}(u)] = u$$

where $F_X^{-1}(u) = inf\{x: F_X(x) \geq u\}$.

The probability integral transform allows one to convert marginal distributions into Uniform (0,1) distributions. With information about the marginal CDFs, the probability integral transform also allows for simulation of random variables that follow other distributions (Casella & Berger, p. 55).

Determining the Appropriate Copula Model

Chi-Square Test of Independence

Before moving forward with the use of a copula model, one must first determine if the variables being modeled exhibit a relationship on each other, or if they are independent. If two random variables are independent, then using a copula model is an inefficient method. A definition of statistical independence is provided below.

Definition: Two events A and B are independent if $P(A \cap B) = P(A)P(B)$

In terms of random variables X and Y, with densities $f_x(x)$and $f_Y(y)$, if X and Y are independent, then their joint density can be written as:

$$f_{X,Y}(x, y) = f_x(x)f_Y(y)$$

In short, statistical independence says that knowing information about X does not affect what is known about Y and vice versa. There are different ways to test for statistical independence. One of the most common is to consider a Chi-square test of independence. In this test, consider the bivariate case, with two random variables, X and Y. Then the following hypotheses exist:

H_0: X and Y are independent

H_1: X and Y are not independent

To reject H_0, one needs to find the test statistic χ_ν^2, where ν represents the degrees of freedom.

χ_ν^2 is calculated by the formula:

$$\chi_\nu^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed value in the ith partition, E_i is the expected value of points in the ith partition, and ν = (number of rows - 1) * (number of columns - 1).

Once this test statistic has been calculated, one can use a Chi-square distribution Table to test the χ_ν^2 value against a critical value based on the number of degrees of freedom and a predetermined level of significance (α). If χ_ν^2 is larger than the critical value, H_0 (independence) is rejected. If it is rejected, then it is assumed X and Y exhibit some relationship with each other, even if that relationship cannot be determined easily. It is these latter situations that may lend themselves to the use of a copula model.

Consider several observations, $\boldsymbol{x} = x_1, \cdots, x_n$ and $\boldsymbol{y} = y_1, \cdots, y_n$, from two random variables X and Y with CDFs $F_x(X)$ and $F_Y(Y)$. Consider a percentile plot of bivariate paired data using the points $(F_x(x_i), F_Y(y_i))$ for $i = 1, \cdots, n$. One can divide this grid into several partitions. Consider the case where 100 partitions are used, where each partition is 10% by 10% of the data. If X and

Y were independent random variables, one would expect each partition to hold 1% of the paired observations. If H_0 is rejected, then a significant number of the partitions contain either significantly more or less than 1% of the data. Figure 1 provides a graphical representation of H_0 (independence) within a percentile plot.

Figure 1: A graphical representation of H_0 (independence) in the chi-square test of independence. 1% of the data is expected in each square.

Left-Right Function

If independence is rejected, tail dependence is a valuable tool in determining which copula should be used (Brehm et al). The Left-Right (LR) concentration function is a graph that uses the ideas from upper and lower tail dependence to show the tail strengths for a certain copula model given a certain value of Kendall's Tau (Brehm et al). This function is defined on $(0 < z < 1)$ and is segmented into its left and right components in a piecewise manner.

Definition: Given two random variables X and Y with CDFs $F_X(X)$ and $F_Y(Y)$. The LR concentration function is defined by:

$$LR(x) = \begin{cases} P[F_X(X) < z \mid F_Y(Y) < z], & 0 < z \leq 0.5 \\ P[F_X(X) > z \mid F_Y(Y) > z], & 0.5 < z < 1 \end{cases}$$

The LR function should be interpreted as the conditional probability that $F_X(X)$ is small given $F_Y(Y)$ is small or $F_X(X)$ is large given $F_Y(Y)$ is large.

In the context of this project, given certain data, one can calculate Kendall's Tau and create the LR concentration function. Then, the copula that best fits that data can be determined (Brehm et al). Figure 2 is a reference graph that can be used to make this determination for a Kendall's Tau value of 0.35. In it, the differences in tail behavior between various copulas can be seen.

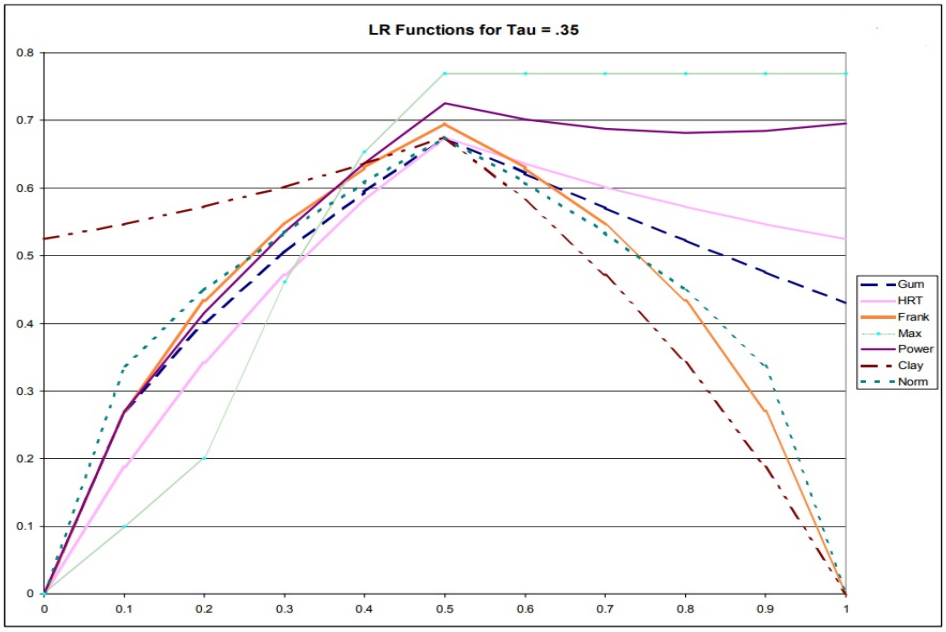

Figure 2: LR functions for a variety of copula with Tau = 0.35 (Brehm et al, p.122)

Types of Copulas

Just like there are well known univariate distributions, there are many well-known copulas (Klugman et al). Many of the parameters that these copulas take can be estimated as a function of Kendall's Tau. Thus, having empirical data and information on the LR concentration function allow a modeler to choose a copula and then to solve for the given parameters in the copula. One basic, yet sometimes trivial, copula is the Independence Copula.

Independence Copula

The Independence Copula is the copula that results from a dependency structure in which each individual variable is independent of one another. The Independence Copula is one of several Archimedean Copulas, another type of copula, and the special case of the Gaussian Copula with a correlation matrix equal to the identity matrix (Klugman et al). For the bivariate case, let X_1 and X_2 be independent random variables. The corresponding independence copula is

$$C(u_1, u_2) = P(U_1 \leq u_1, U_2 \leq u_2)$$

$$C(u_1, u_2) = P(U_1 \leq u_1) * P(U_2 \leq u_2)$$

$$C(u_1, u_2) = u_1 u_2$$

Heavy Right Tail Copula

Another important copula that can be useful in insurance applications is the Heavy Right Tail (HRT) copula. The HRT copula allows for less correlation in the lower tail (smaller losses) and higher correlation in the upper tail (large losses, which is typically important in actuarial applications) (Brehm et al). The formula for this copula is:

$$C(u, v) = u + v - 1 + [(1-u)^{-1/\alpha} + (1-v)^{-1/\alpha} - 1]^{-\alpha}$$

where $\alpha > 0$ can be found using the closed form of Kendall's Tau for this copula:

$$\tau_K(\alpha) = \frac{1}{2\alpha + 1}$$

Normal Copula

One final copula that proves extremely useful is the Normal copula, specifically because it is very easy to simulate (Brehm et al). It is lighter in the upper right tail than the HRT copula, but still assigns more correlation between variables in the upper right and lower left tails (Brehm et al). The formula for this copula is:

$$C(u, v) = \Phi(p(u), p(v)|a)$$

where $\Phi(x, y|a)$ is the bivariate standard normal CDF with correlation a. The closed form of Kendall's Tau for the normal copula is:

$$\tau_K(a) = \frac{2arcsin(a)}{\pi}$$

Marginal Distribution Determination and Simulation

As mentioned before, a copula can help to couple two non-independent sets of data, regardless of the distribution of each. Multivariate models that can only account for variables that follow the same distribution (i.e., the bivariate normal case) are often seen, but copulas enable one to use variables that follow different distributions with different sets of parameters. In many applications, it is important to know these true marginal distributions to analyze how effective the copula is in terms of prediction. Thus, a knowledge of important and common univariate distributions is fundamental. Some of the ones of most interest in the actuarial field are the Gamma distribution, the Weibull distribution, the Lognormal distribution, the Pareto distribution, and the Burr distribution. Each of these has their own number of distinct parameters that impact the shape and scale of the distribution, and these can be estimated from empirical data in a variety of ways. The most common method for parameter estimation is Maximum Likelihood Estimation (MLE), but there are several other approaches that may be used (Brehm et al).

Once the parameters are estimated, one must then evaluate the fit of the resulting distribution to the empirical data. One important visual indicator is the p-p plot. The p-p plot is a graph that helps to determine whether an empirical data set fits a given probability distribution. This plot, an example of which in shown in Figure 3, compares the empirical CDFs of the data with that of the assumed true CDFs. If the plot of these two distributions is approximately linear, it indicates that the assumed true distribution gives a reasonably good fit to the data. In areas where it is below this line, then the theoretical CDF is assigning more probability than what was seen in the empirical data, and in areas where it is above the line, it is assigning less probability than what was seen in the empirical data. Once the estimated distribution is validated, the simulation of new data can begin.

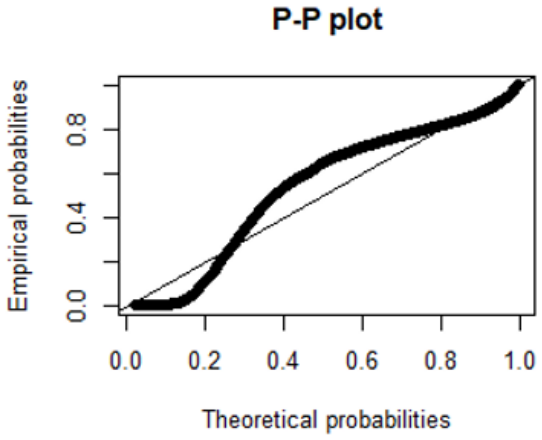

Figure 3: An example p-p plot generated in R

Simulation is an important part of any modeling exercise. It allows one to observe how their model may react to certain initial conditions, how different models compare to empirical data,

and how changing parameters will affect the outcomes of different models. In the field of probability and statistics, modeling is a crucial step in testing how close a model's results are compared to what was observed empirically. Simulation is also an extremely helpful tool in generating new, random data that can be evaluated.

Suppose one would like to simulate n pairs of observations that come from random variables X and Y, where X and Y follow a certain copula. The steps to simulate are:

1. Determine the distributions and parameters.
2. Generate two sets of Uniform (0,1) observations.
3. Invert the copula formula to solve for and generate a new, correlated vector.
4. Apply F_X^{-1} and F_Y^{-1} to the set of uniform observations to create vectors X and Y.

To demonstrate this process, consider the Heavy Right Tail (HRT) copula.

1. Determine the distribution and parameters.

Consider two empirical distributions, X and Y, that are determined to follow the distributions:

$$X \sim \text{Gamma } (\alpha, \beta)$$

$$Y \sim \text{Uniform } (a, b)$$

For this copula, Kendall's Tau is a function of parameter α:

$$\tau_K(\alpha) = \frac{1}{2\alpha + 1}$$

Thus, it can be found that:

$$\alpha = \frac{1}{2\tau} - \frac{1}{2}$$

2. Generate two sets of Uniform (0,1) observations.

In this case, because copulas are based on two vectors of data, one would generate two random vectors: **u** ~ Uniform (0,1) and **p** ~ Uniform (0,1).

3. Invert the copula formula to solve for and generate a vector for the other parameter.

Next, parameter α can be used to solve for v using the conditional distribution, $C_1(u, v)$, the derivative of the closed form of the copula (Brehm et al). For the HRT copula, these formulas are:

$$C_1(\boldsymbol{u}, \boldsymbol{v}) = 1 - \left[(1-\boldsymbol{u})^{-\frac{1}{\alpha}} + (1-\boldsymbol{v})^{-\frac{1}{\alpha}} - 1\right]^{-\alpha-1} (1-\boldsymbol{u})^{-1-\frac{1}{\alpha}}$$

$$C_1^{-1}(\boldsymbol{u},\boldsymbol{p}) = \boldsymbol{v} = -1 - [\left(\frac{(1-\boldsymbol{p})}{(1-\boldsymbol{u})^{-1-\frac{1}{\alpha}}}\right)^{\frac{-1}{\alpha}-1} + 1 - (1-\boldsymbol{u})^{-\frac{1}{\alpha}}]^{-\alpha}$$

Using this equation, vector **v** can be generated. Vectors **u** and **p** are independent, whereas vectors **u** and **v** will exhibit a similar Kendall's Tau and tail dependence to the empirical data. In fact, these vectors are simulated percentiles of random variables X and Y.

4. Apply $F_X^{-1}$and F_Y^{-1} to the sets of observations to create vector X.

Using the inverse CDFs for each of the distributions that were estimated in step one, one can use **u** and **v** to create new vectors **x** and **y** in the following way.

$$\boldsymbol{x} = F^{-1}_{X\,|\,\alpha,\beta}(u_i), \qquad i = 1,\cdots,n$$

$$\boldsymbol{y} = F^{-1}_{Y\,|\,a,b}(v_i), \qquad i = 1,\cdots,n$$

X and Y follow the marginal distributions that were estimated from the empirical data. The data simulated using **u** and **v** maintains the same association that is indicated by the copula.

Our Project

For this project, the team worked with The Hanover Insurance Group to obtain and analyze insurance data and provide them with insights into how they can integrate copula modeling into their models. The Hanover is a property-casualty insurance company located in Worcester, MA, and they provided data from their Commercial Lines division. Within Commercial Lines, The Hanover highlighted three lines of business for analysis: Business Owner's Policy (BOP), Commercial Package Policy (CPP), and Workers' Compensation. More information on these business lines and their associated data can be found in Appendix A.

The data contains a collection of claims from accident years 2007 through 2020, including the trended ultimate losses and trended ultimate defense and containment costs (DCCE) associated with each claim. The team was asked to investigate whether a copula could be introduced into their modeling process for losses and DCCE for large loss claims. Currently, The Hanover aggregates large losses, losses over $1,000,000, and their associated DCCE, and models the sum as one severity distribution. This is mostly done for practicality reasons. The team was tasked with determining the copula that offered the most appropriate fit to the empirical losses and DCCE, creating a model that used that copula, and analyzing the results in comparison to the results obtained by using their traditional severity modeling method.

Methodology

Objectives

The goals of this project were to investigate the basic mechanics of using copulas to model random phenomena, determine when it was appropriate to use a copula for modeling, and to apply these findings to real world data. The team created a copula model for Worker's Compensation claims and compared its results to claims modeled by techniques currently used by The Hanover Insurance Group. In particular, the team investigated whether there was a benefit to using a copula when modeling large losses. There were four main phases of the project:

Phase 1: Data Cleaning and Exploratory Data Analysis

Phase 2: Copula Selection

Phase 3: Distribution and Parameter Estimation

Phase 4: Simulation and Evaluation

Phase 1: Data Cleaning and Exploratory Data Analysis

The goal of this phase was to get a more thorough understanding of the available data. This helped the team determine where it was appropriate to use a copula, and to calculate important empirical measures that would be used later in the simulation phase. The raw data provided by The Hanover contained values for losses and DCCE that were not relevant to the goals of this project. These values were the losses and DCCE amounts of $0 or less. Thus, the team removed all zero or negative dollar amounts for both losses and DCCE. This was done on a pairwise basis, so if either the loss or DCCE of a claim had a negative or zero value, the entire pair was removed. Once these losses and DCCE were removed from the data, the team moved forward conducting preliminary data analysis on three of The Hanover's commercial lines of business (Workers' Compensation, BOP, and CPP).

In the preliminary data analysis, the team calculated important descriptive statistics such as the mean, standard deviation, quartiles, and counts for losses, DCCE, and the sum of losses and DCCE. The goal of this was to determine which data sets were most appropriate for the use of a copula model. The team calculated Pearson's correlation coefficient and Kendall's Tau between losses and DCCE for each line of business. This analysis revealed that Workers' Compensation had a higher value of Kendall's Tau between losses and DCCE than any of the other lines. Additionally, Worker's Compensation had a much higher claim count, which made it a better candidate than other lines for copula modeling. Thus, the team chose to move forward with Worker's Compensation data.

Phase 2: Copula Selection

The next step was to select the copula that offered the best fit to the empirical data. First, the team created a percentile plot of losses and DCCE to visualize any obvious patterns in the Workers' Compensation data. To create this percentile plot, the team used the index function in Microsoft Excel to determine where each loss and DCCE value was ranked in the data. To determine the percentile, the value of the index function was divided by the total count, returning the percent of observations less than or equal to the observation being tested. These percentiles of losses and DCCE were kept as pairs to keep the information from the same claim together. These values were then plotted against each other with DCCE percentile on the Y axis and loss percentile on the X axis.

After visual inspection, the team conducted a Chi-square test of independence on loss percentiles and DCCE percentiles. The null hypothesis, that the data were independent, was rejected, suggesting claims in Workers' Compensation had losses and DCCE that were dependent upon each other. Had the team found the data to be independent, it would not be an efficient or fruitful exercise to continue searching for a copula model because it is well known how to combine two independent distributions.

Using the previously calculated value for Kendall's Tau, the team then developed the LR function from the empirical Workers' Compensation data. The team referenced the LR function displayed in figure 2, and the tails of the empirical LR function indicated the HRT copula may be appropriate for this data. However, the reference graph had a different Kendall's Tau value than the empirical data, so the team created a theoretical LR function for the HRT copula that exhibited the same Kendall's Tau seen in the data. To create this theoretical LR function, the team used the LR function procedure on two sets of 100,000 observations which followed the correlation structure of the HRT copula at the empirical Kendall's Tau value (how these vectors were created will be discussed in Phase 4). Admittedly, this was not an exact theoretical graph, but by using a very large sample size of 100,000, this function converged to the true theoretical values. From these graphs, the team selected the HRT copula for modeling Workers' Compensation losses and DCCE.

Phase 3: Distribution and Parameter Estimation

In phase three, the team determined the approximate marginal distributions of the empirical losses, DCCE, and their sum. Using MLE, the approximate parameters for five common severity distributions were identified. These distributions included the Weibull, Burr, Lognormal, Gamma, and Pareto distributions, most of which can be accessed through the "actuar" package in R. For each distribution, the team estimated the parameters using the fitdist() function from the "fitdistrplus" package, which has MLE as one of the options for parameter estimation. In total, 15 distributions were produced: five for losses, five for DCCE, and five for the sum of losses and DCCE. Once each distribution was estimated, the team used the resulting p-p plot to decide which distribution and parameters were most appropriate for modeling the marginal distribution of losses, DCCE, and their sum. The p-p plot that showed the closest relationship between the empirical and theoretical distributions were the Burr distribution for losses, the Pareto distribution for DCCE, and the Burr distribution for the sum of losses and DCCE.

The Burr distribution is commonly used to fit insurance claim sizes and it has three parameters (a, b, s). Its CDF is given by:

$$F_X(x) = 1 - [1 + (\frac{x}{s})^b]^{-a}$$

The inverse CDF, F_X^{-1}, for this distribution is:

$$F_X^{-1}(x) \ = s\ [(1-u)^{-\frac{1}{a}} - 1]^{\frac{1}{b}}$$

The Pareto distribution is used to model data that has heavy tails, and it has two parameters (a, s). Its CDF is given by:

$$F_X(x) = 1 - (\frac{s}{x})^a$$

The inverse CDF, F_X^{-1}, for this distribution is:

$$F_X^{-1}(x) \ = \frac{s}{(1-u)^{1/a}}$$

Phase 4: Simulation and Evaluation

Following the steps provided in the background, the team simulated values from the above distributions. There were two models created: losses and DCCE simulated from separate, marginal distributions which, with the aid of a copula, were modeled as a joint distribution, and the sum of losses and DCCE modeled as one distribution. The steps for the simulations in each model are given here.

Copula Model:

1. Using R, the team generated two sets of 100,000 uniform random variables:

   p ~ Uniform (0,1)
   u ~ Uniform (0,1)

2. The next step was to calculate parameter "α" using the value of τ_K. For this HRT copula:

$$\tau_K = 0.513$$
$$\alpha = \frac{1}{2\tau} - \frac{1}{2} = \frac{1}{2(0.513)} - \frac{1}{2} = 0.475$$

3. Then, the team calculated $\boldsymbol{v} = C_1^{-1}(\boldsymbol{u}, \boldsymbol{p})$. This would connect randomly generated percentiles with the copula. As noted in the background, for the HRT copula:

$$C_1^{-1}(\boldsymbol{u}, \boldsymbol{p}) = \boldsymbol{v} = -1 - [\left(\frac{(1-\boldsymbol{p})}{(1-\boldsymbol{u})^{-1-\frac{1}{\alpha}}}\right)^{\frac{-1}{\alpha}-1} + 1 - (1-\boldsymbol{u})^{-\frac{1}{\alpha}}]^{-\alpha}$$

   The team used the above formula to generate a new vector **v** that was based on vectors **u** and **p**. This left the team with **u** and **v**, where each u_i were observations from a Uniform (0,1) distribution, and each v_i an observation between 0 and 1. **u** and **v** should exhibit the same Kendall's Tau, and tail dependence structure as the original empirical data.

4. The penultimate step was to apply the inverse CDF of the appropriate distribution to these vectors. For losses, this was:

$$F_{Losses}^{-1}(\boldsymbol{u} \mid a, b, s) = s\,[(1-\boldsymbol{u})^{-\frac{1}{a}} - 1]^{\frac{1}{b}}$$

   This was done in R using the qburr() function.

   For DCCE, the inverse CDF was:

$$F_{DCCE}^{-1}(\boldsymbol{v} \mid a, s) = \frac{s}{(1-\boldsymbol{u})^{1/a}}$$

   This was done in R using the qpareto() function.

5. Finally, the team took these new observations for losses and DCCE and summed them together to get an estimate for the total loss and DCCE attributed to each "claim".

$$Total\ Losses\ and\ DCCE\ = F^{-1}_{Losses}(\boldsymbol{u}) + F^{-1}_{DCCE}(\boldsymbol{v})$$

Non-Copula Model:

The model for the sum of losses and DCCE was simpler because it did not involve a copula.

1. The team started by generating a different random sample than the ones used in the copula model, still having 100,000 observations:

   $$\mathbf{w} \sim \text{Uniform } (0,1)$$

2. Then the team used the inverse CDF of the proposed marginal distribution to come up with observations:

   $$F^{-1}_{Losses\ and\ DCCE}(\boldsymbol{w} \mid a, b, s)\ = s\ [(1 - \boldsymbol{w})^{-\frac{1}{a}} - 1]^{\frac{1}{b}}$$

   This was calculated in R using the qburr() function.

3. Since these values already accounted for losses and DCCE, the team did not need to sum anything.

   $$Total\ Losses\ and\ DCCE\ = F^{-1}_{Losses\ and\ DCCE}(\boldsymbol{w})$$

After simulating observations for both models, the team noticed that several values for the sum of losses and DCCE were much larger than any values observed in the empirical data. This caused both models to seem much more heavily right skewed than the empirical data, affecting the variance and means of the data sets. So, the team removed all simulated values of total loss and DCCE above ten million dollars, because the largest sum of loss and DCCE in the empirical data was less than $9,000,000. Additionally, it was infeasible to keep these values in the data analysis if they would never occur in real life, as some of these losses were in the billions and trillions. For the copula model, the team removed 754 values, and for the non-copula model, the team removed 875 values.

Because the focus of this project was the feasibility of copula modeling on large losses, the team zoomed in on the top ten percent and top five percent of the remaining results for each sample. The team calculated the 90th and 95th percentiles of total losses and DCCE for each model, the tail value at risk for the 90th and 95th percentiles, along with the standard deviation and coefficients of variation for values above the 90th and 95th percentiles. These served as summary statistics for the highest values of losses and DCCE and gave the team an easy way to compare the two modeling methods in the upper right tail.

Results and Discussion

Using the steps outlined in the methodology, the team made two models that demonstrated how a copula could be used to model losses and DCCE in a specific line of business. This section will reveal some key findings from this analysis, discuss how the results from the copula model compare to the current modeling methods at The Hanover, and explain some future steps that may improve the model.

Workers' Compensation and Copulas

After cleaning the data of all negative and zero values, it was clear that Workers' Compensation had significantly more claims than the other lines of business. Additionally, the value for Kendall's Tau between losses and DCCE within Workers' Compensation was much higher than the values in other business lines, as seen in Table 4. Although the linear correlation wasn't the largest, Worker's Compensation had the greatest Kendall's Tau value, which captured the non-linear relationships between losses and DCCE better than the Pearson correlation coefficient.

Business Line	Claim Count (non-zero losses and DCCE)	r	τ_k
BOP Liability	5,229	0.266	0.385
BOP Property	4,433	0.428	0.368
CPP Liability	14,525	0.425	0.420
CPP Property	5,114	0.623	0.447
Worker's Comp.	69,728	0.451	0.513

Table 4: Claim Count, Pearson Correlation, and Kendall's Tau by Business Line

Due to time constraints, the team wanted to focus on only one line of business for this project. The abundance of data and strong dependency between losses and DCCE in Workers' Compensation made that line of business the best candidate to use a copula to model for.

Selecting the HRT Copula

In the percentile plot, displayed in figure 3, one can see the data shows heavy concentration in the upper right and lower left tails. This showed the team that there could be a relationship between losses and DCCE, even if it was not linear. The Chi-square test of independence calculated a p-value of approximately zero, which led the team to reject the null hypothesis that losses and DCCE were independent of each other.

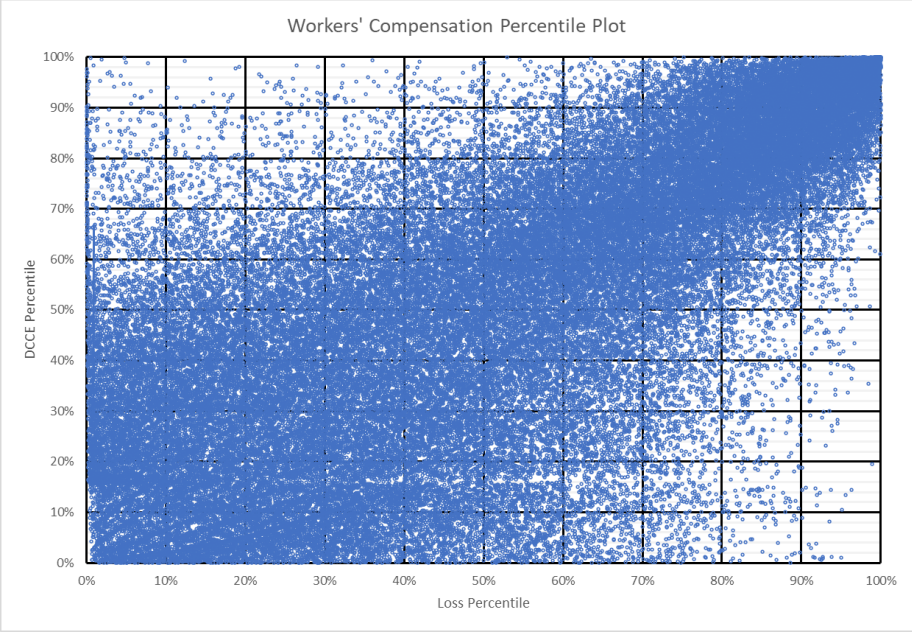

Figure 4: The percentile plot of Workers' Compensation losses and DCCE

The LR graph, which can be seen in figure 4, was created from the empirical Worker's Compensation data, having a Kendall's Tau of 0.513.

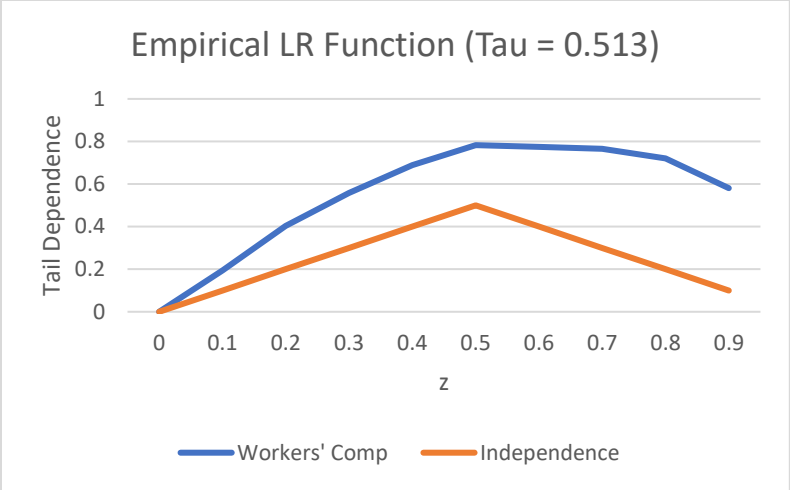

Figure 5: LR Function for Workers' Compensation

This LR function had a shape most similar to that of the HRT copula, based on the reference LR function for various copulas with a Kendall's Tau of 0.35 shown in figure 2. After creating a theoretical LR function for the HRT copula with Kendall's Tau of 0.513, seen in figure 5**,** it was clear the HRT copula was the best choice for modeling Worker's Compensation claims. It is important to note that for the empirical data, the tail of the LR function starts to decrease for a z value of 0.8 more significantly than what is seen in the theoretical LR function. This could suggest there is a better copula to describe the relationship between loss and DCCE. But, based on the copula models known to the team, the HRT copula had the most similar LR function to the empirical data.

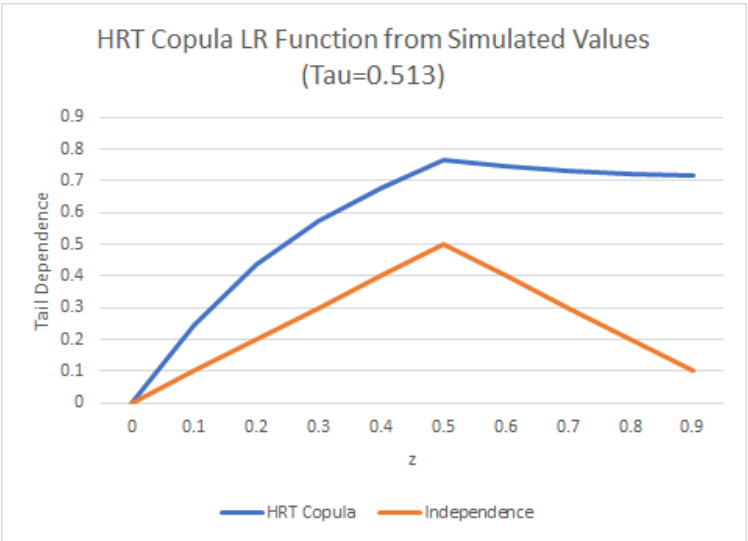

Figure 6: HRT Copula LR Function from Simulated Data

Estimated Distributions and Parameters

After performing MLE on the empirical data in R, the team determined that the following distributions were most appropriate. The CDFs for these distributions can be found in the methodology.

Losses:
Burr Distribution - Shape Parameter 1 (a) = 0.2045, Shape Parameter 2 (b) = 2.3350, Rate Parameter (s) = 0.0029

DCCE:
Pareto Distribution - Shape Parameter (a) = 0.4750, Scale Parameter (s) = 16.6585

Sum of Losses and DCCE:
Burr Distribution - Shape Parameter 1 (a) = 0.1745, Shape Parameter 2 (b) = 2.6459, Rate Parameter (s) = 0.0029

Copula Model vs. Non-Copula Model

A main goal of the project was to determine how the copula model generated total losses and DCCE of a claim, versus how the traditional, single-distribution method would generate the total losses and DCCE of a claim. To do this, the team considered the top 10% and the top 5% of the distributions of each phenomenon, and calculated the means, standard deviations, and coefficients of variation for each model, which can be seen in Table 5.

	90th Percentile		95th Percentile	
	Copula Model	Non-Copula Model	Copula Model	Non-Copula Model
VaR	$41,528	$42,261	$157,761	$155,580
TVaR	$703,326	$717,308	$1,324,868	$1,352,989
St Dev	$1,461,655	$1,496,667	$1,870,281	$1,915,474
CV	2.078	2.087	1.412	1.416

Table 5: Results for the copula model vs. the non-copula model

These statistics do not differ significantly between the models at either the 90^{th} or the 95^{th} percentile. Thus, the models offer very similar results, and the copula model does not provide any significant improvement over the traditional modeling method. One likely explanation is that the size of the losses dominates the size of any associated DCCE within a claim. Because losses are often so much larger than the DCCE of a claim in Workers' Compensation, modeling losses and DCCE as a sum is very similar to just modeling losses by themselves. This is also apparent in the similarities between the distribution and parameters of losses and the sum of losses and DCCE. However, if there were lines of business where DCCE makes up a more significant amount of the total claim, than a copula model may yield significantly different results than the traditional modeling method.

Potential Future Steps

1. Incorporate spliced distributions
2. Investigate statistical tests to evaluate goodness of fit
3. Apply copula modeling in other lines of business

1. Incorporate spliced distributions

As can be seen in Table 5, the estimates for the TVaR and the standard deviation of the top simulated claims are much larger than what is seen in the empirical data. This is mainly due to the limitations the team faced when estimating the distributions for losses, DCCE, and their sum. Using MLE, the team was not able to set any limitations for how high the losses and DCCE could be because the distributions chosen had infinite domains. Without a maximum, very large values of losses and DCCE were generated. Even though values over 10 million dollars were removed from the simulation data, the model was still over-predicting large losses in general. If this model were to be improved in the future, the team would propose exploring a spliced distribution for losses. If a spliced distribution were used, one may be able to control the probability that a large loss occurs. With this decreased probability of large losses, the mean and standard deviation of the models at the upper percentiles would decrease and be a better fit with the empirical data.

2. Investigate statistical tests to evaluate goodness of fit

Another step that the team would have liked to achieve, given more time, would be to investigate statistical tests that could be used to obtain evidence to support the modeling choices they made. Most of the justification given in this report for copula selection and distribution estimation was based on plots (LR Functions, p-p plots, etc.) and other visual indicators. The team discussed several possibilities of different tests that could be used to evaluate the choice in copula and distributions, and test whether the differences in the model were statistically significant, but ultimately ran out of time. If this project were to be continued, it is recommended that the new team investigate some possible statistics or tests that would provide a more robust adjudication of the modeling selections and estimates.

3. Apply copula modeling in other lines of business

Lastly, the team believes that if this model was applied to a line of business where the DCCE made up a more significant amount of the total claim, than it may offer more insight into the use of copula modeling at The Hanover. Although this project confirms that The Hanover's current large loss modeling methods are reasonable in Workers' Compensation, copulas are still worth investigating in other business lines.

Conclusion

Copulas are a statistical tool that can help guide insurance companies in extreme economic situations. By taking advantage of different measures of dependence, it is possible to study the underlying data structure between multiple random variables. This information allows a modeler to determine an appropriate copula model for a set of empirical data. Once this copula is chosen and estimates of the marginal distributions have been made, a modeler can create a joint probability model for several random variables.

This project focused on connecting losses and DCCE of insurance claims in large loss scenarios. The team was able to create a copula model for use with Workers' Compensation data, finding that the HRT copula was the most appropriate. Through parameter estimation for the marginal distributions, the team was able to simulate losses and DCCE that followed the data structure originally observed. While this model did not yield significantly different results than traditional modeling methods used by The Hanover, it did confirm that The Hanover's current method for modeling Worker's Compensation claims is valid.

References

Brehm, P. J., Gluck, S. M., Kreps, R. E., Major, J. A., Mango, D. F., Shaw, R., Venter, G. G., White, S. B., & Witcraft, S. E. (2007). *Enterprise risk analysis for property & liability insurance companies: A practical guide to standard models and emerging solutions*. (G. R. Perry, Ed.). Guy Carpenter & Co.

Casella, G., & Berger, R. L. (2002). *Statistical inference*. Duxbury.

Klugman, S. A., Panjer, H. H., & Willmot, G. E. (2013). *Loss models: further topics*. John Wiley & Sons, Inc.

Appendix A: The Hanover Commercial Lines

Commercial lines insurance provides property and liability coverage to businesses, rather than to individuals. Property insurance covers damage to buildings, cars, and other physical products, while liability insurance protects businesses from claims that result from an injury to a person or damage to another person's property. Within commercial lines at The Hanover, there are three main lines of business, Business Owner's Policy (BOP), Commercial Package Policy (CPP), and Workers' Compensation. BOP provides property and liability insurance for smaller businesses, while CPP provides similar coverages to larger businesses that have more risks associated with them. BOP would cover a lawsuit stemming from an accident in a mom-and-pop shop, but CPP might cover the loss of a retail freight truck that is involved in an accident. Workers' Compensation protects employers from claims associated with injuries to their employees that occur while they are on the job. This type of insurance would cover medical expenses, legal fees, and lost wages that may result from work-related incidents.

Appendix B: Fitting Large Losses with the Normal Copula

In insurance, it is typically of more interest to analyze the right tail of the distribution because that is where the insurance company pays large dollar amounts. For this reason, the team considered cutting the data so that it would solely focus on large losses. To cut the data, the team determined where large and attritional losses, the smaller loss values that wouldn't be considered, would be separated. The team considered a few different thresholds for large loss amounts: $1,000,000, $750,000, $500,000, $250,000, and $100,000. The team decided to split data based on loss amount since the losses are the more significant portion of the total claim amount. So, even if a claim had a large sum for its loss and DCCE, it was dropped if the loss did not reach the specified threshold.

When splitting the data at these loss thresholds, the counts of remaining claims dropped significantly. More notably, the values for Kendall's Tau dropped closer and closer to zero, with some even becoming negative, as seen in Table 6.

Claims Included	Number of Claims	Kendall's Tau (τ_k)
All Claims	69,728	0.513
Claims With Losses Above $100,000	3,184	0.245
Claims With Losses Above $250,000	900	0.150
Claims With Losses Above $500,000	251	0.127
Claims With Losses Above $750,000	120	-0.027
Claims With Losses Above $1,000,000	66	-0.062

Table 6: Number of Claims and Kendall's Tau at different loss thresholds

Having small and negative values of Kendall's Tau indicates the data do not have a strong positive relationship with each other. This was surprising since the initial data analysis implied there was a relationship between losses and DCCE, particularly in the right tails, which is what this data had been limited to. Thus, the only threshold where copula modeling made sense was when investigating separating claim sizes at $100,000.

The team then developed an LR function for the empirical data above this threshold, with a Kendall's Tau of 0.25, which can be seen in figure 6**.** This graph was compared to the reference graphs provided in the background, and it was decided that the Normal Copula was the best fit.

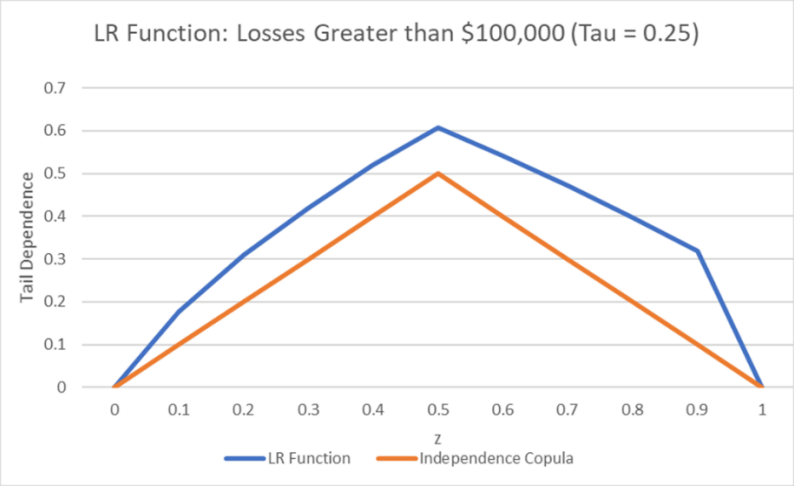

Figure 7: LR Function for Workers' Compensation Losses over $100,000

Once the team had selected the normal copula, they were able to simulate values using the process outlined in the background and methodology but with the normal copula. From these simulated values, a theoretical LR function was developed. The team expected the theoretical LR function to diverge from the LR function of independence, which would indicate some sort of dependence between the variables. However, the LR function the team created from the normal copula did not diverge from independence, as seen in figure 7. This indicated that the normal copula was a poor representation of the empirical data. Additionally, the team did not have sufficient information to determine which model would be a better fit for the data, so the team abandoned the idea of using a normal copula to model this data.

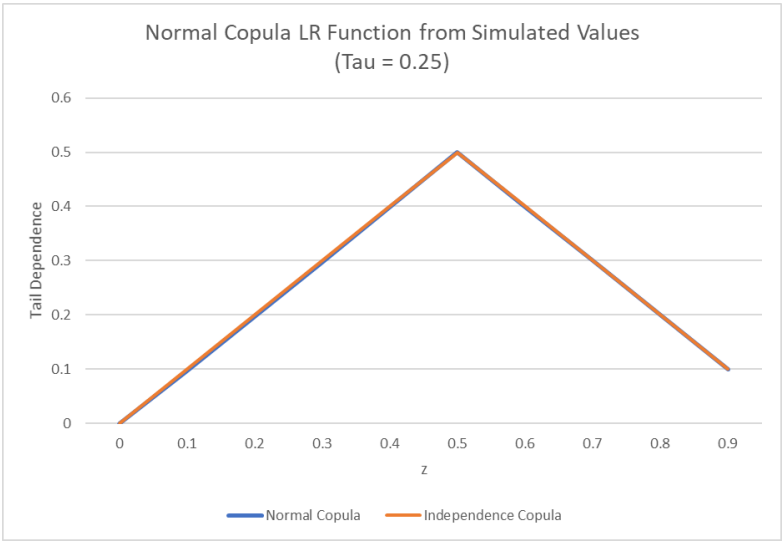

Figure 8: LR Function for Simulated Values of the Normal Copula

The team believes this did not work out well because the Kendall's Tau value was too low, almost to the point where the normal copula was graphing two independent phenomena. As mentioned previously, splitting the data in any way resulted in the loss of significant amounts of data. There were also many values of DCCE, particularly some very large values, that were taken out because of the sole focus on large losses. By removing so many of these data points, the data structure was changed heavily and there was no longer a strong relationship in the right tails of the distribution. When looking to separate attritional losses from large losses, the choice of copula is made to represent the underlying structure and dependency measures of all the data. If the goal is to separate attritional and large losses while still preserving these values, one should consider using spliced or mixed distributions to better account for different patterns at different times in the data, rather than just removing data altogether.