USING COPULAE TO ANALYSE COMPETING RISKS

Modelling time-to-closure of mortgages due to default or prepayment

James Gray, Louis Kearney, Hugh Langan & Niall O'Donovan

May 2025

Abstract - In this paper we examine the use of copulae, statistical tools which allow us to model the dependency structures of related random variables, from the perspective of the competing risks framework. Initially we formally define copulae and describe related theory, giving examples of some potential applications. We then apply these concepts to a real-world example, the Freddie-Mac Single Family Loan-Level dataset. We analyse the joint distribution of two time-to-event variables; loan prepayment and loan default. We employ copulae to better model these variables, avoiding obscured data caused by only one event being able to occur in each datapoint. We incorporate the knowledge that the obscured time-to-event is greater than that of the observed, defining the copula accordingly.

1 Introduction

Time-to-event datasets generally take the form of a time variable T, alongside an event-type label I [1]. One way of viewing T is as a combination of cause-specific time-to-events, $(T_1, T_2 \dots)$, where T_i is the time until event i [2]. In many scenarios, the occurrence of one event prevents the subsequent observation of other potential events [3]. This is prevalent in medical datasets looking at time-to-death of a subject due to different causes, or in financial datasets where events represent the closure of a contract due to different reasons. In these cases the event which did occur has *masked* the other events, *obscuring* the other time-to-events. We can define the overall time-to-event $T = \min\{T_i;\ i \in \mathbb{N}\}$, or for a case with two possible events; $T = \min(T_1, T_2)$.

It is clear from this definition of time-to-event variables that there is some level of dependency between them. We can effectively model them by fitting marginal distributions, with copulae to describe the dependency structure [4]. Before defining in greater detail the concept of a copula, and the field of competing risks, we will first review some basic concepts necessary below.

1.1 Joint distributions

When working with two random variables, X and Y, the joint probability distribution describes the likelihood of any possible pair of outcomes. The probability of observing a specific pair of outcomes is written as:

$$P(X = x,\ Y = y) \text{ or } P(X, Y)$$

This concept can be extended naturally to n random variables. The joint probability of $X_1, X_2, \dots, X_n$ is denoted as:

$$P(X_1 = x_1,\ X_2 = x_2,\ \dots,\ X_n = x_n)$$

$$\text{or } P(X_1, X_2, \dots, X_n)$$

The joint distribution can also be fully described by its cumulative distribution function (CDF):

$$F_{X_1}(x_1) = P(X_1 \leq x_1, \dots, X_n \leq x_n)$$

$$= \int \cdots \int f(x_1, \dots, x_n) dx_1 \dots dx_n$$

The joint distribution captures not only the behaviour of individual variables but also how they are related. From the joint distribution, one can compute the marginal distributions for any subset of variables. This relationship will be explored further in the discussion of Sklar's Theorem.

1.2 Marginal distributions

A marginal distribution refers to the probability distribution of a single random variable, or a subset of variables, within a joint probability distribution. It represents the behaviour of one variable irrespective of the others.

To obtain the marginal distribution from the joint distribution, we integrate out the variables that are not of interest. This process, called marginalisation, involves summing or integrating over all possible values of the unwanted variables.

Consider a set of random variables $X_1, X_2, \dots, X_n$. The marginal distribution of X_1 is given by:

$$f_{X_1}(x_1) = P(X_1 \leq x_1)$$

$$= \int \cdots \int f(x_1, x_2, \dots, x_n) dx_2 \dots dx_n$$

This integral effectively "collapses" the joint distribution into a distribution for just X_1, by accumulating probability across all values of the other variables.

Similarly, the marginal CDF of X_1 can be defined as:

$$F_{X_1}(x_1) = P(X_1 \leq x_1)$$

$$= \int \cdots \int F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) dx_2 \ldots dx_n$$

It is more common (and simpler) however to express this using the joint density function if it exists. In that case:

$$F_{X_1}(x_1) = \int_{-\infty}^{x_1} f_{X_1}(x_1)\, dx_1$$

$$= \int_{-\infty}^{x_1} \left(\int \cdots \int f(x_1', x_2, \ldots, x_n) dx_2 \ldots dx_n \right) dx_1'$$

With this approach we first integrate out all the other variables to get the marginal density $f_{X_1}(x_1)$, then integrate that function from $-\infty$ to x_1 to get the marginal CDF.

1.3 Grade of a random variable

Given any random variable X, we can transform it using its own cumulative distribution function (CDF), F_X, to define a new variable:

$$U = F_X(X)$$

This transformation is called the grade of X. The resulting variable U follows a uniform distribution on the interval $[0, 1]$, regardless of the original distribution of X:

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

This transformation is critical because it standardizes variables regardless of their original distributions. That is, no matter how complex or skewed the original distribution of X was, the grade U will always be uniformly distributed [5].

2 Copulas

Let $X = (X_1, \ldots, X_n)$ be a vector of random variables with a joint probability density function f_X. As discussed earlier, we can obtain the marginal distributions of each X_i by integrating f_X. Once we have the marginal CDFs F_{X_i}, we can apply them to each X_i to form the vector of grades [4], [5]:

$$u_i = F_{X_i}(X_i)$$

The copula of X, denoted f_U, is defined as the joint distribution of these grades:

$$f_U = F_{U_1,\ldots,U_n}$$

This distribution is defined on the unit cube $[0, 1]^n$, and captures all the dependence structure among the components of X, independent of their marginals. If the variables X_i are independent, then their grades U_i will also be independent, and the copula reduces to the uniform distribution on $[0, 1]^n$, otherwise known as the independence copula.

To build a formal definition of the copula, we rely on **Sklar's Theorem**, which states that any multivariate distribution can be expressed in terms of its marginals and a copula function that encodes their dependency [4]. We know that there must be some copula C that represents the dependency structure between the components of X, so we can say:

$$F(x_1, \ldots, x_n)$$

$$= C(F_{X_1}(x_1), \ldots, F_{X_n}(x_n)) = C(u_1, \ldots, u_n)$$

If F is continuous, then the copula C will be unique, and we can recover it via:

$$C(u_1, \ldots, u_n) = F(F_{X_1}^{-1}(u_1), \ldots, F_{X_n}^{-1}(u_n))$$

where F is the joint CDF, and $F_{X_1}, \ldots, F_{X_n}$ are the marginals for each component of random vector X. This means we can reconstruct the copula by evaluating the joint CDF at the inverse marginals. This process transforms uniform values back to their original scales before evaluating F.

The above equation is obtained by transformation of the CDF of U, f_U:

$$F_U(U) = P(U_1 \leq u_1, \ldots, U_n \leq u_n)$$

$$= P(F_{X_1}(X_1) \leq u_1, \ldots, F_{X_n}(X_n) \leq u_n)$$

$$= P(X_1 \leq F_{X_1}^{-1}(u_1), \ldots, X_n \leq F_{X_n}^{-1}(u_n))$$

$$= F_X(F_{X_1}^{-1}(u_1), \ldots, F_{X_n}^{-1}(u_n))$$

If the joint distribution has a density, we can define the copula density c by differentiating the copula function with respect to all its arguments:

$$c(u_1, \ldots, u_n) = \frac{\partial^n C(u_1, \ldots, u_n)}{\partial u_1, \ldots, u_n}$$

Using this, we can express the joint probability density function f_X of the original variables in terms of the copula density and the marginal densities [6]:

$$f_X(x_1, \ldots, x_n) = c(F_{X_1}(x_1), \ldots, F_{X_n}(x_n)) \cdot \prod_{i=1}^{n} f_{X_i}(x_i)$$

This formula cleanly decouples the distribution into two components:

- The marginal densities f_{X_i}, which describe how each variable behaves on its own.
- The copula density c, which models have the variables interact or co-vary.

This decomposition has important implications:

- It allows for flexible modelling: We can choose arbitrary marginal distributions for each variable (e.g., normal, exponential, Pareto) and pair them with a suitable copula that reflects the desired dependence (e.g., Gaussian, Clayton, Gumbel).
- It facilitates modular design in statistics and machine learning. Instead of fitting a full multivariate distribution, we can model marginals and dependence separately, then recombine them [7].
- It's particularly powerful in domains where extremal dependence matters, such as:
  - **Finance**: tail dependence between asset returns [8].
  - **Insurance**: joint risk of large claims [9].
  - **Hydrology**: co-occurrence of extreme rainfall and river discharge [10].
  - **Healthcare**: competing risks of clinical outcomes [11].

2.1 Example: Sklar's Theorem in finance

Let's consider a finance-related example using Sklar's Theorem to model the relationship between two variables: stock market returns (X) and interest rates (Y). In financial modelling, understanding the dependency between economic indicators and asset returns is vital for portfolio optimisation, risk management, and stress testing [12]. Copulas provide a flexible way to model this relationship. The process can be outlined as follows:

Step 1: Analyse marginal distributions

First, examine the marginal distribution of stock market returns (X). This might involve calculating return distributions for different market indices (e.g., S&P 500, NASDAQ) or individual stocks, looking at volatility, skewness, and fat tails in returns.

Then, analyse the marginal distribution of interest rates (Y), such as short-term treasury yields or central bank policy rates. You could study the distribution over different macroeconomic cycles (e.g., recession, expansion), noting historical volatility and rate clustering.

Step 2: Identify a copula to model dependency

Thanks to Sklar's Theorem, we know there exists a copula that captures how interest rates and market returns co-move, regardless of their individual distributions. The goal is to choose a copula that best fits this observed dependence.

This involves examining how returns tend to behave in environments of rising or falling interest rates. For example, do stock returns exhibit stronger left-tail dependence (i.e., simultaneous downturns) when rates spike unexpectedly? In this case we might select the Clayton copula.

By applying Sklar's Theorem, financial analysts can separate the modelling of marginal behaviours (e.g., heavy-tailed returns, bounded interest rates) from their dependency structure (e.g., asymmetric correlations). This leads to more robust risk models, improved portfolio stress testing, and better understanding of systemic risk in financial markets.

3 Competing risks

Competing risks occur when the occurrence of one type of event affects the likelihood—or precludes the possibility—of other event types. These situations arise frequently in time-to-event (survival) analysis. For example, in medical studies, death due to cancer may prevent the observation of death due to heart disease. In sports, if a team concedes a goal, it might increase the chance of committing a foul (a dependent competing risk), or make other outcomes irrelevant (mutually exclusive) [3].

In statistical terms, we consider datasets consisting of pairs $\langle T, I \rangle$, where:

- $T^{(i)}$ is the time-to-event for subject i.
- $I^{(i)}$ is the indicator for the type of event (i.e. which risk occurred)

A common modelling approach assumes each subject has a latent time-to-failure for each risk, denoted $T_1^{(i)}, \ldots, T_K^{(i)}$. The actual observed time is:

$$T^{(i)} = \min(T_1^{(i)}, \ldots, T_K^{(i)})$$

Only the event corresponding to the minimum time is observed - failures from other risks are censored. We call the unobserved times conceptual failure times. To model them, we assume:

$$T_k \sim F_{T_k}(u_k)$$

where

$$(u_1, \ldots, u_n) \sim C$$

Here, F_{T_k} is the marginal distribution for cause k, and C is a copula that encodes the dependence structure between the risks via their associated uniform random variables. This model assumes that dependency structure between risks is described fully by the copula C, which is completely independent from the nature of the risks' marginals [2]. This separability is an assumption that follows from Sklar's Theorem. In practice you must still choose marginals that fit the data."

3.1 Copulae & competing risks

Recall that a copula is a multivariate distribution function defined on the unit cube $[0,1]^n$, used to model the dependency structure between variables. By applying the CDF transformation $U = F_X(X)$, any continuous random variable can be mapped to a uniform variable on $[0,1]$. A copula C, combined with marginal CDFs, $F_{X_1}, \ldots, F_{X_n}$ allows us to construct a full multivariate distribution.

This is particularly valuable in competing risks, where we often suspect non-trivial dependencies between the conceptual times to failure. For example, failure due to one disease might biologically correlate with failure due to another — dependencies that standard multivariate models might miss or misrepresent.

A canonical financial analogy is joint stock collapse: two companies may have very different marginal risk profiles, but their failure times may be highly correlated during a systemic market crash. Copulas allow us to model such rare but critical dependencies accurately [13].

3.2 Challenges in competing risks data

One major difficulty in this setting is that we never observe the full vector of conceptual failure times — we only observe the minimum [3]. For example, if T_p and T_q represent times to failure from two causes p and q, we never observe both unless they are equal (an event of zero probability in continuous time). Instead, we observe:

$$T = \min(T_p, T_q)$$

$$I = \operatorname{argmin}(T_p, T_q)$$

This masking makes it difficult to estimate the true dependence between T_p and T_q [14]. For instance, observing $I = p$ implies $T_p < T_q$, but does not reveal how close the values were, or how likely one would have been to occur soon after the other.

Despite this, copulas can still be used to infer and estimate the dependency structure. Under certain assumptions (e.g., known marginals or parametric families), we can fit a copula model to observed outcomes and estimate parameters that reflect the latent correlation structure between risks.

This is where copulas shine: they decouple the shape of each failure time distribution from the structure of their dependency, allowing for a more flexible and interpretable model of complex risk behaviour. With correctly selected copulas, we can infer the missing data and draw meaningful insights.

4 Types of copulae

The simplest kind of dependency structure is no dependence at all - i.e., when the variables are statistically independent. This situation is modelled using the independence copula, defined as:

$$C(u_1, \ldots, u_n) = \prod_{i=1}^{n} u_i$$

This copula corresponds to a multivariate uniform distribution over $[0,1]^n$ and reflects the idea that knowing the value of one variable tells you nothing about the others.

Other copulas give describe more complex relationships between random variables.

4.1 Archimedean copulae

A widely used and versatile class of copulas is the Archimedean family. These copulas are popular because they are easy to construct (requiring only one parameter), yet still flexible enough to capture various dependency structures, including asymmetric tail dependence.

An Archimedean copula is defined by a generator function ψ, a continuous, strictly decreasing convex function from $[0,1]$ to $[0,\infty]$, satisfying $\psi(1) = 0$. Its inverse ψ^-1 maps $[0,\infty]$ back to $[0,1]$. The copula formula for n variables is:

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

Different choices of ψ correspond to different copulae within the family. Three common examples include:

- **Clayton copula:** $\psi(t) = t^{-\theta} - 1$
- **Gumbel copula:** $\psi(t) = (-\log(t))^{-\theta}$
- **Frank copula:** $\psi(t) = -\log(\frac{e^{-\theta t}-1}{e^{-\theta}-1})$

In each case, a single parameter θ is used to control tail dependence, the likelihood of extreme values occurring together. Tail dependence refers to the strength of association in the extreme ends of the distribution. There are two types:

1. **Upper tail dependence:** Simultaneous large (positive) values across variables.
2. **Lower tail dependence:** Simultaneous small (negative/extreme low) values.

The Clayton and Gumbel copulas are used to capture lower and upper tail dependencies respectively [4]. The Frank copula is used in scenarios where central correlation is more relevant and heavy tails are not anticipated.

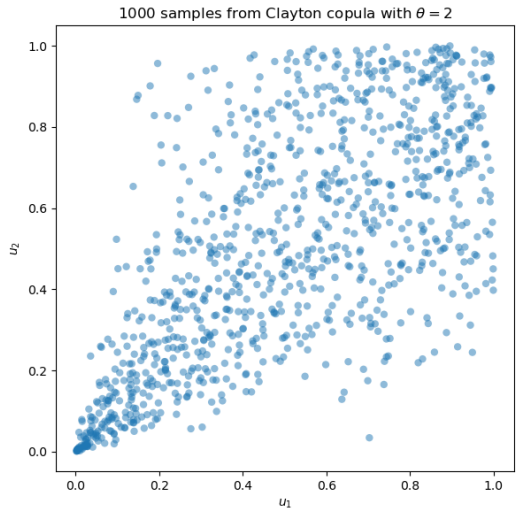

Figure 1: Samples taken randomly from the Clayton copula, which has a lower tail dependence

4.2 Elliptical copulae

Another major family of copulas is the elliptical copulas, which includes the well-known Gaussian (normal) and t-Student copulas. These copulas derive from elliptical distributions [6], a generalization of the multivariate normal distribution.

Elliptical copulas are especially useful when the data naturally follows a symmetric structure and when correlation is the main form of dependence. However, they come with trade-offs: while Gaussian copulas are easy to use and interpret, they cannot capture tail dependence, which can be a major drawback in risk-sensitive applications. In contrast, t-copulas can model tail dependence effectively and are therefore more robust in the presence of joint extreme events.

4.2.1 Multivariate elliptical distributions

To understand elliptical copulas, we first need to grasp multivariate elliptical distributions. A random vector $X \in \mathbb{R}^n$ is said to follow an elliptical distribution if it satisfies one of the following:

1. **Density function form:**

$$f_X(x) = k \cdot g((x-\mu)^T \Sigma^{-1}(x-\mu))$$

Here μ is the center of the distribution, Σ is a positive definite dispersion matrix, $g(\cdot)$ is a scalar function that controls shape, and k is a normalising constant.

2. **Characteristic function form:**

$$\phi_X(t) = \psi(t^T \cdot \Sigma \cdot t)$$

where ψ is a scalar function and Σ is again the dispersion matrix.

These definitions generalize the multivariate normal distribution, which is a special case when g is an exponential function. Elliptical distributions can accommodate heavy tails (e.g., multivariate t-distributions), making them better suited to real-world data that deviates from normality.

4.2.2 Elliptical copula: formal definition

Given a random vector $X = (X_1, \ldots, X_n)$ following an elliptical distribution with CDF F_X, we define the elliptical copula using Sklar's Theorem [6]:

$$C(u_1, \ldots, u_n) = F_X(F_{X_1}(u_1), \ldots F_{X_n}(u_n))$$

Here, F_{X_i} are the marginal CDFs, which do not necessarily have to be elliptical. Unlike Archimedean copulas, which are defined using generator functions, elliptical copulas are defined implicitly through the CDF of the joint elliptical distribution.

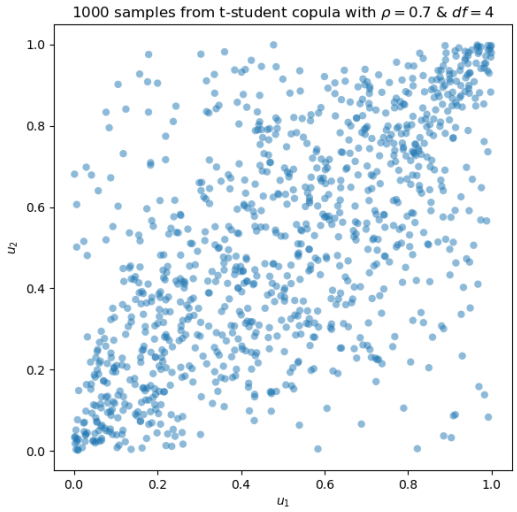

Figure 2: Samples taken randomly from the t-Student copula, which has a symmetric tail dependence

4.2.3 Gaussian copula

$$C(u_1, \ldots, u_n; \Sigma) = \Phi_\Sigma(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_n))$$

- Φ_Σ is the CDF of a multivariate normal with correlation matrix Σ.
- Φ^{-1} is the inverse of the standard normal CDF.

While Gaussian copulas are easy to work with and popular in many real-world applications, they cannot model joint extremes, i.e., they have zero tail dependence [6]. This means they may underestimate joint risk in extreme values. Famously, Gaussian copulas misunderstood extreme market shifts during the 2008 financial crisis [15].

4.2.4 t-Copula

$$C(u_1, \ldots, u_n; \Sigma, v) = T_{\Sigma,v}(T_v^{-1}(u_1), \ldots, T_v^{-1}(u_n)$$

- $T_{\Sigma,v}$ is the CDF of a multivariate t-distribution with correlation matrix Σ and degrees of freedom v.
- T_v^{-1} is the inverse of the univariate t-distribution.

The t-copula exhibits symmetric tail dependence in both upper and lower extremes, making it suitable for modelling joint crashes or surges. The level of tail dependence is controlled by the degrees of freedom, with lower values creating heavier tails [6].

5 Simulated analysis

The real dataset analysed in section 6 is a time-to-closure dataset detailing the number of days a mortgage remained active until the holder either prepaid or defaulted. We will attempt to model the dependency between these two competing risks using a Clayton copula. After justifying its use, we will explore some techniques important for analysis on simulated datasets.

5.1 Justification For Use Of Clayton Copula

The Clayton copula offers an advantage over the other copulae in the given context because of its strong lower tail dependence and asymmetric dependence structure. Lower tail dependence means that the Clayton copula assigns extra probability to joint occurrences of extremely small values of the variables beyond what would be expected under independence or symmetric dependence models.

This captures the intuition that if borrowers who pay early or default have similar risk profiles, both being more reactive to things such as rate movements, job losses, or personal events. If one termination event happening early, the other is also likely to occur early (this is the case even though for a single loan only the first event is observed; lower tail dependence implies a tendency for both latent default and prepayment times to be short). This property is desirable in the mortgage context. Early defaults and prepayments often arise from common shocks or borrower characteristics in the loan's initial period. Loans subject to such early shocks have an elevated risk of either default or prepayment in a short time frame.

The Clayton copula's concentration of mass in the lower tail reflects this by coupling the risks during the initial years. It increases the likelihood that both the default time and prepayment time are simultaneously low, aligning with scenarios where an early shock could have led to either happening. This reveals itself as a higher incidence of early mortgage terminations, a feature that copulas lacking tail dependence would miss but the Clayton can easily capture.

Another importat feature of the Clayton copula is its asymmetric dependence. While there is a strong association in the lower tail, there is much weaker dependence in the upper tail. This asymmetric property is useful to mortgage life-cycle behaviour.

While many loans that terminate do so relatively early, those that survive the initial high-risk period tend to behave more differently later on. In other words, if neither default nor prepayment has occurred in the early years, the timing of very late-term defaults or prepayments may be caused by more independent or borrower-specific factors that do not necessarily coincide.

The Clayton copula accommodates this by not imposing excessive dependence in the upper tail. It allows the joint survival of both risks to be nearly independent in the tail, meaning a loan that hasn't defaulted or been prepaid for a long duration doesn't automatically imply a high likelihood of the other event occurring at the same late stage. A symmetric copula would not distinguish between early and late dependence. it would either underplay early correlations or over-impose correlation in the later stage.

The fit of the Clayton copula over the others is most evident when compared to the Gaussian copula. In the context of mortgage risk, a Gaussian copula would significantly underestimate the probability that a loan susceptible to early default is also susceptible to early prepayment. It cannot increase the joint probability of both events occurring at short durations beyond what an average correlation implies. This shortcoming has had fallouts in the past. It was the failure of Gaussian copulas to capture joint extreme events that resulted in the mispricing of credit products in the 2008 financial crisis.

5.2 Comparing true and masked distributions of mortgage time-to-closure

Suppose we are working with a mortgage dataset that tracks a large number of loans over time. Each loan eventually either defaults or is fully prepaid. These are competing risks: once one occurs, the other is no longer possible. Our goal is to estimate the time-to-event distributions for each of these two outcomes.

Let:

- T_1: the time until default
- T_2: the time until pre-payment

In this setting, we only observe the first event that occurs for each loan. That is, for loan i we observe:

$$T^{(i)} = \min(T_1^{(i)}, T_2^{(i)})$$

and an indicator variable:

$$I^{(i)} = \begin{cases} 1 \text{ (default)} & T_1^{(i)} < T_2^{(i)} \\ 2 \text{ (pre-payment)} & T_1^{(i)} > T_2^{(i)} \end{cases}$$

It is guaranteed that only one event has occurred and therefore $T_1 \neq T_2$.

Each entry in our dataset is therefore a pair $(T^{(i)}, I^{(i)})$, indicating the observed time and the event type.

To simulate such a dataset, suppose both default times T_1 and prepayment times T_2 are drawn from independent Weibull distributions. That is, we assume independence between these competing risks in this simulation. In practice, the assumption could be relaxed using a copula, but we begin with this simpler case.

We can simulate the data by first drawing uniform random variables $U_1, U_2 \sim C_{\text{ind}}$, and then applying the inverse CDFs of the respective Weibull distributions:

$$T_1 = F_{T_1}^{-1}(U_1),\ T_2 = F_{T_2}^{-1}(U_2)$$

The true distributions of T_1 and T_2 (before censoring) are known to us, but in a real-world scenario, we would not observe both times for any loan.

In this simulation, we generate 1000 such loan records. For each loan, only the minimum of T_1 and T_2 is observed, along with an indicator of whether the event was a default or a prepayment. This leads to censoring:

- If $T_1 < T_2$, pre-payment time is censored.
- If $T_2 < T_1$, default time is censored.

In particular, the observed data will under-represent longer event times, since those are more likely to be censored by an earlier competing event. This censoring leads to underestimation of the mean event times if not properly accounted for.

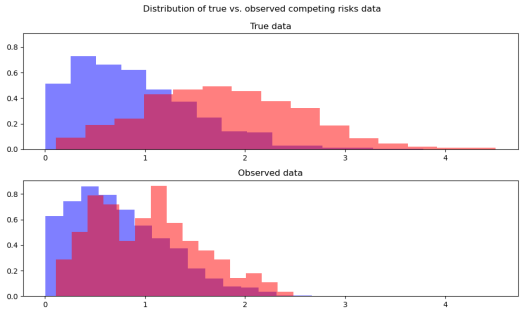

Figure 3: The distribution of the obscured data is biased, with the Weibull with the lower mean better represented

To correct this censoring, we construct a likelihood function over all observed data, including both [16]:

- Exact event times for the observed event.
- The fact that the unobserved event occurred after the observed time.

For example, if our datapoint has $t = 50$ and $i = 1$ (pre-payment occurred), then we can say:

$$T_2 = 50,\ T_1 > 50$$

This datapoint would contribute to the likelihood function the term [3]:

$$l(t, i) = P(T_2 = 50 \wedge T_1 > 50)$$

The overall likelihood function over the dataset $\langle T, I \rangle$ will be defined:

$$\mathcal{L}(\langle T, I \rangle; \lambda, k) = \prod_{(t,i)} l(t, i)$$

where:

$$l(t, i) = \begin{cases} P(T_1 = t) \cdot P(T_2 > t) & \text{if } i = 1 \text{ (default)} \\ P(T_2 = t) \cdot P(T_1 > t) & \text{if } i = 2 \text{ (pre-payment)} \end{cases}$$

Here λ and k are the scale and shape of the distribution.

By maximizing this likelihood function over the whole dataset, we can estimate the underlying distribution parameters more accurately than if we used only the uncensored data. The key insight is that every data point contains information, even if it's only interval-censored (e.g., we know one time is greater than another).

	k_1	λ_1	k_2	λ_2
True	1.5	1.0	2.5	2.0
Naive	1.5286	0.8292	2.0367	1.1510
Likelihood	1.4613	0.9918	2.3428	2.0708

Table 1: The likelihood function approach sees improved results over fitting just the observed data

The results seen in Table 5.2 show that the marginals' parameters can be recovered with great accuracy through use of the likelihood approach, certainly much better than using only the observed data. Encouraging to see is that the likelihood approach's improvements are especially strong on distribution two, which was obscured much more heavily due to its higher mean.

5.3 Fitting a Clayton copula to censored data

Assuming the ability to find the correct parameters for the marginal distributions of the two time-to-closure variables, lets now attempt to fit a Clayton copula to the sample data.

Our main approach will be to use a custom likelihood-based approach, similarly to how we fit the Weibull distributions in the previous simulated test. For comparison, we also evaluate a baseline approach using purely random imputation. In both cases, we will generate censored test data using a fixed value for θ, then try to recover it using the data only.

Again we simulate mortgage data T_1 and T_2 representing time to default and time to pre-payment. Both follow known Weibull distributions, and we assume that

their dependence structure follows a Clayton copula with known parameter θ. The goal is to recover this θ using only the observed data:

$$T = \min(T_1, T_2),\ I = \begin{cases} 1 & T_1 < T_2 \\ 2 & T_2 < T_1 \end{cases}$$

To simulate the dataset we generate 1000 uniform values from a Clayton copula with $\theta = 3.0$, then transform each pair to time values using the appropriate Weibull CDF. The minimum of these two values and its appropriate label are taken for each (t, i) in $\langle T, I \rangle$, with the others discarded. We repeat the trial with 10 randomly generated datasets to avoid any one-off variance.

Random imputation

This method generates plausible full data from the observed dataset by randomly imputing the missing event time for each observation:

- For each (t, i), the unobserved variable (either T_1 or T_2) is missing.
- Based on the indicator I, we draw the missing copula uniform variable $u_k \sim \text{Unif}(u_j, 1)$, where $j \neq k$, and $F_{T_j}(T) = u_j$, giving us a completed pair (u_j, u_k).

We do know that imputation introduces uncertainty and does not make full use of the known structure of censoring, possibly reducing precision.

Likelihood method

This method uses a tailored likelihood that accounts for the partial information available in censored observations.

The likelihood for each datapoint is constructed as:

$$l(t, i) = \begin{cases} P(U_2 > F_{T_2}(T) \mid U_1 = F_{T_1}(T)) & \text{if } i = 1 \\ P(U_1 > F_{T_1}(T) \mid U_2 = F_{T_2}(T)) & \text{if } i = 2 \end{cases}$$

For the case $i = 1$, we can define:

$$P(U_2 > F_{T_2}(T) \mid U_1 = F_{T_1}(T))$$

$$= \frac{\partial C_\theta}{\partial u_1}(u_1, u_2)$$

$$= 1 - (u_1^{-\theta} + u_2^{-\theta} - 1)^{-(1+1/\theta)} \cdot u_1^{-(1+\theta)}$$

with similar for $i = 2$ but by finding $\frac{\partial C_\theta}{\partial u_2}(u_1, u_2)$.

The overall likelihood function is then given as:

$$\mathcal{L}(\theta; \langle T, I \rangle) = \prod_{(t,i)} l(\theta; t, i)$$

which can be used to find θ using an maximum likelihood estimate (MLE) approach. In this case Python library scikit-learn's 'minimize' function was used to minimise the negative log likelihood.

	True	**Naive**	**Likelihood**
$\hat{\theta} \pm std.dev.$	3.0	1.574 ± 0.114	2.481 ± 0.151

Table 2: The likelihood function approach gives far greater accuracy over the the random imputation method

The results seen in Table 5.3 show significant improvement using the likelihood approach over a naive random imputation approach. Both methods underestimate the strength of the dependence between the two variables, but likelihood approach is consistently closer. It should be acknowledged that the likelihood method does not magically recover unavailable information from the obscured data, and so it would be unreasonable to expect a perfect prediction. Instead it uses as much information as possible from the nature of the inputted data to give the best estimate possible.

5.4 Estimating marginals & copula together

So far, we've explored how to estimate the copula parameter θ assuming perfect knowledge of the marginals, and how to estimate the marginal parameters assuming θ is known. In this section, we tackle the more realistic and challenging scenario of estimating both the dependence structure and the marginal distributions simultaneously, using only the censored observations $\langle T, I \rangle$.

We once again simulate a dataset containing 1000 loans. The datapoints are generated with the same method as in the previous simulation, but in this case we discard all parameters for both Weibull distributions and the Clayton copula, looking to recover them.

A similar likelihood function can be used as in the previous step, just with all of the variables (shape and scale for both marginals, θ for the copula) included in the MLE process. As we are now also concerned about optimising the marginals, we also include the probability of getting the observed time data, using the Weibull distribution's PDF.

$$l(t, i)$$

$$= \begin{cases} P(T = t) \cdot P(U_2 > F_{T_2}(t) \mid U_1 = F_{T_1}(t)) & \text{if } i = 1 \\ P(T = t) \cdot P(U_1 > F_{T_1}(t) \mid U_2 = F_{T_2}(t)) & \text{if } i = 2 \end{cases}$$

We use 'L-BFGS-B' for optimisation, which will find the local optimisation of the likelihood function. For this reason a range of initial values were attempted.

Param	True	Estimated
θ	3.0	2.351 ± 0.286
λ_1	1.0	1.000 ± 0.000
k_1	1.5	1.618 ± 0.032
λ_2	2.0	1.974 ± 0.042
k_2	2.5	3.874 ± 0.456

Table 3: Estimation across the board is reasonably good, especially with regards to the marginals

The results seen in Table 5.4 show that the taken approach does a good job of estimating the overall dataset parameters. The estimation of θ is perhaps the weakest of the parameter estimations, under-estimating the strength of the relationship between the two variables. This under-estimation is likely symptomatic of the lack of information available due to the obscured time-to-event data. The Weibull marginals are fitted very accurately, barring the shape of the second. This reflects the fact that most values sampled from our second marginal will be greater than those from the first, by nature of the original parameters, thus meaning we lose more information here when the data is obscured.

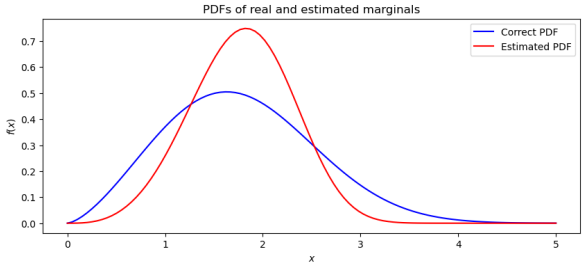

Figure 4: The second estimated marginal is of the correct scale but incorrect shape

6 Real dataset analysis

In this section we will finally analyse the real mortgage dataset; the Freddie Mac Single Family Loan-Level Dataset [17]. This dataset collects a range of performance data on American mortgages between 1999 and 2024. The selected subset was the `sample_2024` data, a randomly selected subset of 50,000 loans. Of this subset, we are concerned with loans that either defaulted or were prepaid, of which there are 125 and 16,897 respectively.

6.1 Data processing

For this use-case, we are only concerned with two columns from the dataset; the 'zero balance code', the 'zero balance effective data', and the 'first payment date'. These first two columns columns describe for what reason the balance of an individual loan reaches zero and at what date this occurs. We can combine the zero balance date with the first payment date to get our time-to-event statistic; the number of days since the loan began. This distribution of this data can be seen below.

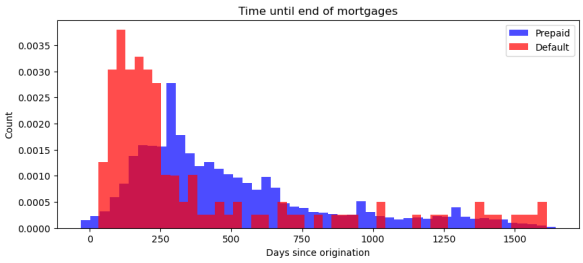

Figure 5: The distribution of the real time-to-event mortgage data

The zero balance code column can take a number of values, only some of which concern us. Code 1 corresponds to loans which were prepaid. Codes 2, 3, 9, and 96 correspond to mortgage defaults for a number of reasons. In this case we combine them into one event.

From this input data we generate a table of time-to-event numbers and event labels for ingestion into our copula fitting code.

6.2 Fitting marginals & Clayton copula

The process of fitting marginals and a copula to this real data should be somewhat similar to how we have done so with the simulated data. We will again assume the marginals follow a Weibull distribution, which is supported by the distributions seen in figure 5.

Some adjustment was required to the log likelihood function due to the large size of the dataset. Instead of multiplying small probabilities, we take the logs of the components and add them as we go. Otherwise, the logic remains the same; using the likelihood function and MLE to find the optimal value for each of the five parameters.

Param	Estimation
θ	3.268
λ_1	556.910
k_1	1.135
λ_2	387.823
k_2	2.394

Table 4: The final results after applying our likelihood approach to fit both marginals and a copula to the time-to-closure mortgage data

6.3 Interpretation of fitted parameters

Key inferences

- *Early-life co-movement.* Clayton $\theta = 3.268$ gives Kendall's $\tau = \theta/(\theta+2) \approx 0.62$ and lower-tail dependence $\lambda_L = 2^{-1/\theta} \approx 0.81$, implying strong clustering of early defaults and pre-payments.

- *Hazard shapes.* Weibull shapes show a sharply rising pre-pay hazard ($k_2 \approx 2.4$) versus a near-flat default hazard ($k_1 \approx 1.1$); expected times are $\mathbb{E}[T_{pre}] \approx 344$ and $\mathbb{E}[T_{def}] \approx 532$ days.
- *Hidden default exposure.* Because many latent defaults are censored by pre-pay, naïve credit-loss forecasts would understate long-horizon risk.
- *Stress-testing implication.* Joint early-termination shocks should be used when valuing MBS tranches.

We can use these parameters and the original competing risks data to recover the unobserved event times. This is done by converting the observed value to its quantile and passing this to the conditional inverse CDF of the fitted Clayton copula. This CDF will return us a quantile to be used in the CDF of the unobserved time-to-closure statistic which is consistent with the dependence structure described by the fitted copula. Doing this for each datapoint gives us the below two-dimensional data.

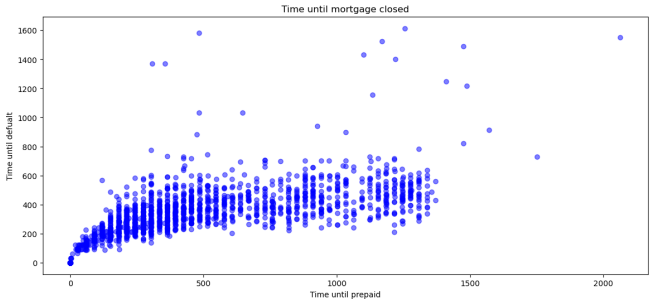

Figure 6: The fitted copula is used to impute the missing time-to-closure data

Similarly, we can sample completely new datapoints from the copula and marginals. Below we have sampled 1250 samples from each marginal, then applied the approach used with the real data, imputing the quantile to be used in the other marginal using the inverse copula CDF.

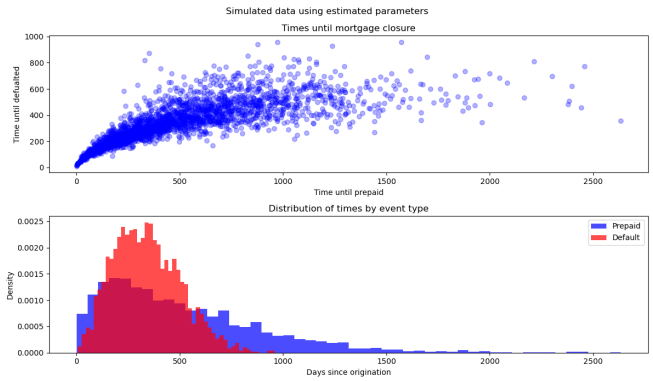

Figure 7: Entirely simulated mortgage closure time data

Interesting to note in the above graph is the differences to the original distribution of real data. Both marginals, but especially that of the time-to-defaults which was more often obscured, have heavier right tails. This is because the majority of recovered datapoints lie in this right tail where the event never occurred first.

7 Conclusion

In this paper, we have given a concise account of the theory of copulas, outlined their application to the analysis of competing risks, explained how to use a likelihood-based approach to deal with the censored data that characterises competing risks, and demonstrated fitting a copula-based model to simulated data.

Finally, we analysed the Freddie Mac Single Family Loan-Level Dataset using a copula-based model. This model provides insight into the dependency structure between loans that defaulted and loans that were prepaid.

Our model highlights the usefulness of the ability of copulas to decouple the dependency structure from the marginal distributions of the variables under investigation. This allowed us much more freedom in modelling by allowing us to choose arbitrary marginal distributions and allowed us to incorporate sensible assumptions about the relationship between defaulting and pre-payment into our model.

The fitted Clayton parameter 3.27 implies Kendall's tau 0.62 and a lower tail-dependence coefficient of about 0.81. This shows that when one latent termination falls in the earliest 1 % of its distribution, the competing risk is roughly 80 % likely to be equally early. When taken alongside the sharply rising pre-payment hazard versus the nearly flat default hazard, this clustering of early terminations highlights why tail-dependent copulas give more realistic stress-test and MBS-valuation results than Gaussian or independence assumptions.

References

[1] J. D. Kalbfleisch and R. L. Prentice, *The Statistical Analysis of Failure Time Data* (Wiley Series in Probability and Statistics), 2nd ed. John Wiley & Sons, 2002.

[2] P. K. Andersen, Ø. Borgan, R. D. Gill, and N. Keiding, *Statistical Models Based on Counting Processes* (Springer Series in Statistics). Springer, 1993.

[3] R. L. Prentice, J. D. Kalbfleisch, A. V. Peterson Jr, N. Flournoy, V. T. Farewell, and N. E. Breslow, "The analysis of failure times in the presence of competing risks," *Biometrics*, vol. 34, no. 4, 1978.

[4] R. B. Nelsen, *An Introduction to Copulas*, 2nd ed. Springer, 2006.

[5] G. Casella and R. L. Berger, *Statistical Inference*. Duxbury, 2002.

[6] H. Joe, *Dependence Modeling with Copulas* (Monographs on Statistics and Applied Probability). CRC Press, 2014.

[7] G. Elidan, "Copula bayesian networks," *Advances in Neural Information Processing Systems*, vol. 23, 2010.

[8] F. P. Cortese, "Tail dependence in financial markets: A dynamic copula approach," *Risks*, vol. 7, no. 4, 2019. [Online]. Available: `https://www.mdpi.com/2227-9091/7/4/116`.

[9] A. Dias, I. Ismail, and A. Zhang, "Copula-based risk aggregation and the significance of reinsurance," *Risks*, vol. 13, no. 3, 2025. [Online]. Available: `https://www.mdpi.com/2227-9091/13/3/44`.

[10] Y. Liu, Y. Liu, Y. Hao, *et al.*, "Probabilistic analysis of extreme discharges and precipitations with a non-parametric copula model," *Water*, vol. 10, no. 7, 2018. [Online]. Available: `https://www.mdpi.com/2073-4441/10/7/823`.

[11] Y. Wei, M. Wojtyś, L. Sorrell, and P. Rowe, "Bivariate copula regression models for semi-competing risks," *Statistical Methods in Medical Research*, vol. 32, no. 10, 2023.

[12] A. Di Clemente and C. Romano, "Calibrating and simulating copula functions in financial applications," *Frontiers in Applied Mathematics and Statistics*, vol. 7, 2021. [Online]. Available: `https://www.frontiersin.org/articles/10.3389/fams.2021.642210/full`.

[13] Y. Liu, P. M. Djuric, Y. S. Kim, S. T. Rachev, and J. Glimm, "Systemic risk modeling with lévy copulas," *Journal of Risk and Financial Management*, vol. 14, no. 6, 2021. DOI: `10.3390/jrfm14060251`. [Online]. Available: `https://www.mdpi.com/1911-8074/14/6/251`.

[14] K. Gogol, C. Krettek, and J. Kopoczek, "Copula-based modeling of competing risks with masked causes of failure," *Mathematics*, vol. 8, no. 7, p. 1117, 2020.

[15] D. MacKenzie and T. Spears, "The gaussian copula and modelling practices in investment banking," *Social Studies of Science*, vol. 44, no. 3, 2014.

[16] J. Beyersmann and M. Schumacher, "Simulating competing risks data in survival analysis," *Statistics in Medicine*, vol. 27, no. 30, 2008.

[17] Freddie Mac. "Single-family loan-level dataset." (2024), [Online]. Available: `https://www.freddiemac.com/research/datasets/sf-loanlevel-dataset`.