Using Copulas to Model Prepayment and Default in Competing Risks Analysis

STU44003: Data Analytics
Trinity College Dublin, The University of Dublin

This paper explores the use of copula models for analysing competing financial risks, with a particular focus on modelling the dependence between loan prepayment and default events. Beginning with a theoretical overview of copula functions and their application within a competing risks framework, we evaluate the performance of different copula families through simulation studies. These simulations demonstrate that copula-based models can effectively recover both marginal and dependence parameters. The methodology is then applied to a real-world dataset of over two million loans from the Lending Club. Marginal distributions are selected using AIC criteria, with a Weibull distribution fitted for prepayment times and a three-component Weibull mixture for defaults, due to evident multimodality. The results highlight the value of flexible marginal modelling and copula-based approaches in capturing complex interdependencies between financial risks.

1. Introduction & Background

Loan default risk refers to the probability that a borrower will default on a loan, namely that a borrower will fail to make full and timely payments of principal and interest in accordance with the terms of the loan they have received. Default risk is one of the two components of credit risk, and generally aims to assess the willingness and capacity of a borrower to service their debt [16]. Lenders naturally face a direct financial loss as a consequence of a borrower defaulting on their loan, though these losses are, in general, accounted for by the interest made on other loans, provided such defaults can be attributed to the individual situations or calamities of the defaulting borrowers, which do not influence the pool of borrowers as a whole. Mass loan defaults, however, have the potential to devastate lenders, borrowers, and members of broader society.

As a consequence of such mass defaults, financial institutions may find themselves restricted to a reduced lending capacity, which has the potential to adversely influence their faculty to operate effectively. This can render such institutions unable to lend funds to other borrowers, and has been shown to amplify the unwillingness of other financial brokers to meet the needs of small borrowers [10]. On a broader scale, an event of mass defaults has the power to threaten the overall stability of the economy, as the dense interconnection of financial institutions tends to propagate such shocks as an event of mass defaults, causing fragility in the wider financial system [1]. Mass loan defaults were a major component of the Financial Crisis of 2008, which led to severe recessions in economies around the globe [5].

It is thus imperative to assess and analyse such loan default risks, appropriately accounting for the various risk factors and competing risks involved. Competing risks refer to the alternative outcomes that prevent the occurence of the primary event in question, the primary event in this case being loan default [2]. Key competing risks in this scenario include prepayment and refinancing of loans, for instance. In this paper, we aim to use copulas to suitably analyse and model loan default risks.

2. Introduction to Competing Risks & Copulas

Copulas in statistics are functions which “couple” multivariate distribution functions with their one-dimensional marginal distribution functions, which motivates the use of the Latin term “copula”, which means a “link, tie or bond” [9]. Multivariate statistical methods are classically based on the multivariate normal distribution, however the multivariate normal assumption is not always appropriate in every application, and is often too strong an assumption to be made in the first place [15]. Copulas offer flexibility in dependency between risks, as opposed to multivariate distribution.

In essence, copulas are mathematical objects which have marginal distributions as well as parameters based on dependence as inputs, and which output joint distributions [6]. A marginal distribution refers to the the probability distribution of a single variable in a multivariate setting. Once the distribution of any given marginal is identified, one can consider the cumulative density function to map all input observations of this marginal to the unit interval, which brings about a transformation to a uniform distribution. Given an arbitrary variable X, we call the transformed random variable obtained by applying the cumulative density function of X to X itself the grade of X, and note that the distribution of this transformed variable is uniform on the unit interval, irrespective of the original distribution of X [8].

Any two given probabilities of distinct events A and B, let us denote them by $P(A)$ and $P(B)$, lie in the unit interval $[0, 1]$, also known as “Probability State Space” [6]. If the probabilities of these two events are not independent, the task of calculating their joint probability becomes much more complex. One cannot simply add these probabilities together, as such a value may not even lie in Probability State Space, and would thus no longer be a probability at all. Copulas allow for such intricate dependencies between variables.

Consider an n-dimensional vector $X = (X_1, \dots, X_n)$ with a general multivariate distribution represented by its probability density function $X \sim f_X$. For each marginal X_i of X, one can calculate the cumulative density function F_{X_i} of X_i and hence find the grade of X_i, denoted $U_i = F_{X_i}(X_i) \sim U_{[0,1]}$. One can then consider the n-dimensional vector $U = (U_1, \dots, U_n)$ of grades. The copula of the distribution f_X is defined as the joint distribution f_U of its grades [8]. It is crucial to note that the entries of U are not independent, thus f_U is not uniform on its domain, which is the unit cube.

A vital result in the study of copulas is Sklar's Theorem, stated formally below.

Theorem 2.1 *Let F be a p-dimensional distribution function with margins $F_1, \dots, F_p$. Then there exists a p-dimensional copula C such that, for all x in the domain of F,*

$$F(x_1, \dots, x_p) = C\{F_1(x_1), \dots, F_p(x_p)\}. \tag{1}$$

If $F_1, \dots, F_p$ are all continuous, the C is unique; otherwise, C is uniquely determined on $RanF_1 \times \dots \times RanF_p$, where $RanH$ is the range of H. Conversely, if C is a p-dimensional copula and $F_1, \dots, F_p$ are distribution functions, then the function F defined by 1 is a p-dimensional distribution function with marginal distributions $F_1, \dots, F_p$ [15].

Sklar's Theorem ensures that a given continuous multivariate distribution constitutes two separate components; univariate margins and multivariate dependence, where one represents the structure of dependence by a copula, and can thus analyse such a structure separately from its margins by studying its copula.

Copulas can capture many different and intricate structures of dependence, such as tail dependence, for instance, which is particularly useful in the study of certain financial assets which may not be strongly correlated day to day, but may correlate subject to strong market movement [11]. Naturally, many different types of copulas exist to account for various different types of dependencies, which will now be outlined in section 3. It is essential to note that although copulas can handle complex dependencies incredibly well, one must be careful to select an appropriate copula to implement, in order to avoid such disastrous repercussions as the Financial Crisis of 2008, which was propagated by the mass adoption of the Gaussian copula in financial models by almost everyone in the realm of finance, from bond investors and Wall Street banks to ratings agencies and regulators [12]. It was not the Gaussian copula itself that brought forth such catastrophic results, but the misinterpretation and over-adoption of it, highlighting the importance of appropriate model selection in data analysis.

3. Types of Copulas

In the following section we will discuss the types of copulas commonly used in quantitative finance, including Archimedean, Gaussian, and Elliptical copulas. Similar definitions and results can be found in [6] and [15]

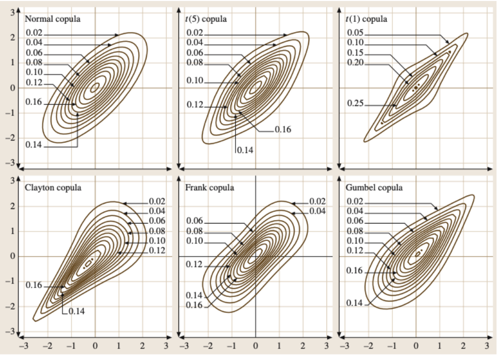


Figure 1. Types of Copulas

Figure 1 retrieved from [15]

3.1. *The Frēcht-Höffding Boundary Copula*

The simplest type of dependence one can consider between multiple variables is independence, namely no dependency at all between variables. The Frēcht-Höffding Boundary Copula can be implemented in the case where the given data is independent. This copula uses the negative logarithm transformation. Two independent probabilities, $P(A)$ and $P(B)$, are transformed using the negative of the logarithmic function, namely

$$-\ln(P(A)) \quad \text{and} \quad -\ln(P(B))$$

This transformation can be inverted by considering the following expression:

$$\exp[-\ln(P(A)) + (-\ln(P(B)))]$$

The independent copula is then given by the function: $C[F(X), F(Y)] = F(X)F(Y)$.

3.2. *Archimedean Copulas*

Archimedean copulas are a family of copulas which are easy to construct and are thus widely used in the case of correlated data. Archimedean copulas make use of generator functions in order to model dependence. A generator function ϕ is a continuous strictly decreasing convex function from $[0, 1]$ to $[0, \infty)$, such that $\phi(1) = 0$. Its pseudo-inverse is given by

$$\phi^{-1}(t) = \begin{cases} \phi^{-1}(t), & 0 \leq t \leq \phi(0) \\ 0, & \phi(0) \leq t < \infty \end{cases}$$

There are several types of Archimedean copulas, including Gumbel, Frank, and Clayton copulas, which will be discussed in greater detail below. These copulas are commonly applied in the fields of finance and engineering, among others. The simple construction of these copulas, however, leads to certain drawbacks in the case of overly complex data. These copulas can be represented as follows,

$$C(u_1, ...u_n) = \varphi^{-1}(\varphi(u_1)...\varphi(u_n)),$$

where $\varphi(u_i)$ are generator functions.

3.2.1. *The Gumbel Copula*

The Gumbel Copula accounts for correlation in the data using the parameter α. The transformation function in question is given by

$$\varphi[F(X)] = [-\ln(F(X))]^{\alpha}$$

The inverse transformation function is given by:

$$\varphi^{-1}[F(X)] = \exp[F(X)]^{1/\alpha}$$

The Gumbel copula is useful in illustrating upper tail dependence. This allows it to effectively model risk under extreme financial conditions. For example, they have been used to quantify co-movements in extreme returns for South African Industrial and Financial Indices, which both exhibit heavy-tailed return distributions [3].

3.2.2. *The Frank Copula*

The transformation function considered in the case of the Frank copula is given by:

$$\varphi[F(X)] = \ln\left(\frac{\exp(-\alpha F(X)) - 1)}{\exp(-\alpha) - 1)}\right)$$

The inverse transformation function is given by:

$$\varphi^{-1}[F(X)] = -\tfrac{1}{\alpha}\ln(1+\exp(-F[X])(\exp(-\alpha)-1)$$

The Frank copula illustrates no tail dependence. This means it is particularly useful for capturing consistent non-extreme co-movements between financial variables. They are also useful in capturing symmetric distributions, which could be valuable in modelling the joint behaviour of various types of financial instruments [7].

3.2.3. *The Clayton Copula*

In the case of the Clayton copula, the transformation function in question is given by:

$$\varphi[F(X)] = \tfrac{1}{\alpha}\left(F(X)^{-\alpha}-1\right)$$

The inverse transformation function is given by:

$$\varphi^{-1}[F(X)] = (1+\alpha F(X))^{\frac{-1}{\alpha}}$$

The Clayton copula can illustrate a lower tail dependence. This characteristic should enable more accurate modelling of scenarios where extreme losses or defaults are likely to occur simultaneously, particularly under adverse economic conditions [4].

3.3. *Gaussian and t-Copulas*

The Gaussian copula is implemented when the given data adheres to a multivariate normal distribution, and the t-copula can be applied to multivariate t-distributions. These copulas are more complex and less explicit than the previously mentioned family of Archimedean copulas. The Gaussian copula in particular is an example of an implicit copula, where any multivariate distribution can be rescaled such that it exhibits uniform marginal distributions.

Limitations of the Gaussian copula can be attributed to its tendency to capture extreme tail dependencies, but the relative simplicity of this copula has led to its wide-spread use. The Gaussian copula is represented as follows,

$$C_R(u_i) = \Phi_R(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_n)),$$

where R is the associated correlation matrix, Φ_R is the joint cumulative density function, and Φ_R is the cumulative density function of a normal variable.

The t-copula extends the Gaussian copula to better control tail behaviour. The t-copula can more effectively model extreme events in comparison to the Gaussian copula, due to the t-copula's use of degrees of freedom to control tail behaviour. The t-copula takes the following form,

$$C_R(u_i; v) = T_{R,v}(F_{t,v}^{-1}(u_1)...F_{t,v}^{-1}(u_n)),$$

where R is the associated correlation matrix, $T_{r,v}$ is the cumulative density function of a multivariate t-distribution, $F_{t,v}$ is a single variable t-distribution, and v represents the degrees of freedom.

4. Methodology

The methodology used in this paper closely follows the framework introduced by Trivedi and Zimmer (2005) [13], who developed a practical approach to modelling joint distributions using copulas. Their work provides the foundation for combining marginal distributions with a copula function to model dependence between variables. In applying this to a competing risks setting, we adopt their structure to model loan prepayment and default times, allowing for flexible dependence between the two risks.

4.1. *Competing Risks Framework*

In the context of modelling loan defaults, we adopt a competing risks framework in which each borrower's time to default is governed by two distinct mechanisms. Specifically, we distinguish between (i) loan prepayment risk, which reflects borrowers paying off their loans early, and (ii) loan default risk, which encompasses financial distress at the borrower level or broader economic downturns affecting the market.

Let T_1 be a random variable representing the time to loan prepayment and let T_2 represent the time to loan default. In this framework, we do not observe both T_1 an T_2 for the same borrower. For each borrower $i \in \{1, \dots, n\}$, we observe only the earliest event:

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

accompanied by an event indicator $I^{(i)}$ defined as

$$I^{(i)} = \begin{cases} 1 & \text{if } T_1^{(i)} < T_2^{(i)}, \\ 2 & \text{if } T_2^{(i)} < T_1^{(i)}. \end{cases}$$

As such, each data point takes the form $(T^{(i)}, I^{(i)})$. Importantly, the assumption of independence between T_1 and T_2 is not imposed, allowing for potential dependence between the failure times.

4.2. *Marginal Survival Distributions*

Consistent with Sklar's Theorem, the copula-based approach requires explicit specification of marginal distributions for each event time. These marginals reflect the temporal risk profiles of prepayment and default, independently of their joint behaviour. We consider two types of marginal specifications depending on the copula model used.

4.2.1. *Case 1: Gamma Marginals*

In the case of the Frank copula, it is assumed that that the time to default, T_1, and T_2 for each competing risk follows a gamma distribution.

Let $T_1 \sim \text{Gamma}(\alpha_1, \beta_1)$, and $T_2 \sim \text{Gamma}(\alpha_2, \beta_2)$, where $\alpha_1, \beta_1 > 0$ denote the respective shape and rate parameters for loan prepayment, and $\alpha_2 > 0$, $\beta_2 > 0$ are the shape and rate parameters for loan default risk, respectively.

The probability density functions for these distributions is given by:

$$f_1(t) = \frac{1}{\Gamma(\alpha_1)} t^{\alpha_1 - 1} e^{-\beta_1 t} \beta_1^{\alpha_1}$$

$$f_2(t) = \frac{1}{\Gamma(\alpha_2)} t^{\alpha_2 - 1} e^{-\beta_2 t} \beta_2^{\alpha_2}$$

$\forall t \geq 0$, where $\Gamma(\alpha)$ denotes the Gamma function.

The cumulative distribution functions for each of the competing risks are:

$$\begin{aligned} F_1(t) = \Pr(T_1 \leq t) = \int_0^t f_1(u)\,du = \gamma(\alpha_1, \beta_1 t) \\ = \int_0^{\beta_1 t} u^{\alpha_1 - 1} e^{-u}\,du, \end{aligned}$$

$$F_2(t) = \Pr(T_2 \leq t) = \gamma(\alpha_2, \beta_2 t) = \int_0^{\beta_2 t} u^{\alpha_2 - 1} e^{-u}\,du,$$

where $\gamma(\alpha, x)$ is the lower incomplete gamma function defined as:

$$\gamma(\alpha, x) = \int_0^x u^{\alpha-1} e^{-u}\, du.$$

The Survival Functions, representing the probability that the event has not occurred by time t, are:

$$S_1(t) = \Pr(T_1 > t) = 1 - F_1(t) = 1 - \gamma(\alpha_1, \beta_1 t),$$

$$S_2(t) = \Pr(T_2 > t) = 1 - F_2(t) = 1 - \gamma(\alpha_2, \beta_2 t).$$

For each observed failure time $T^{(i)} = \min(T_1^{(i)},\ T_2^{(i)})$, the marginal survival functions evaluate to:

$$\Pr(T_1 > T^{(i)}) = S_1(T^{(i)}) = 1 - \frac{\gamma\left(\alpha_1, \beta_1 T^{(i)}\right)}{\Gamma(\alpha_1)},$$
$$\Pr(T_2 > T^{(i)}) = S_2(T^{(i)}) = 1 - \frac{\gamma\left(\alpha_2, \beta_2 T^{(i)}\right)}{\Gamma(\alpha_2)}.$$

The corresponding Cumulative Distribution Values for the transformed uniform variables $U_1^{(i)}$ and $U_2^{(i)}$ are:

$$U_1^{(i)} = F_1(T^{(i)}) = \frac{\gamma\left(\alpha_1, \beta_1 T^{(i)}\right)}{\Gamma(\alpha_1)},$$
$$U_2^{(i)} = F_2(T^{(i)}) = \frac{\gamma\left(\alpha_2, \beta_2 T^{(i)}\right)}{\Gamma(\alpha_2)}.$$

These transformed uniform variables $(U_1^{(i)}, U_2^{(i)}) \in (0,1)^2$ are then passed as arguments to the copula function to specify the joint distribution of the latent event times.

4.2.2. *Case 2: Exponential Marginals*

In the Clayton copula setting, we specify exponential marginals for both competing risks, reflecting memoryless behaviour and constant hazard rates over time. Let $T_1 \sim \text{Exp}(\lambda_1)$ and $T_2 \sim \text{Exp}(\lambda_2)$, where $\lambda_1 > 0$ and $\lambda_2 > 0$ denote the hazard rates associated with loan prepayment and loan default risk, respectively.

The probability density and cumulative distribution functions are given by:

$$f_1(t) = \lambda_1 e^{-\lambda_1 t}, \qquad F_1(t) = \Pr(T_1 \le t) = 1 - e^{-\lambda_1 t},$$
$$f_2(t) = \lambda_2 e^{-\lambda_2 t}, \qquad F_2(t) = \Pr(T_2 \le t) = 1 - e^{-\lambda_2 t},$$

for all $t \ge 0$.

The survival functions, defined as the probability that the event has not occurred by time t, are:

$$S_1(t) = \Pr(T_1 > t) = e^{-\lambda_1 t}, \qquad S_2(t) = \Pr(T_2 > t) = e^{-\lambda_2 t}.$$

At the observed time $T^{(i)}$, the marginal CDF evaluations yield transformed uniform variables:

$$U_1^{(i)} = F_1(T^{(i)}) = 1 - e^{-\lambda_1 T^{(i)}},\ U_2^{(i)} = F_2(T^{(i)}) = 1 - e^{-\lambda_2 T^{(i)}}$$

These uniform marginals $(U_1^{(i)}, U_2^{(i)}) \in (0,1)^2$ serve as inputs to the copula function used to construct the joint distribution of the latent event times (T_1, T_2).

4.3. *Copula Model for Dependence Structure*

Dependence between T_1 and T_2 can be captured by different copulas. In this section we model the dependence structure between the competing risks by using the Frank and Clayton copulas.

4.3.1. *Case 1: Frank Copula*

The Frank copula is also used to model the dependence structure between competing risks, T_1 and T_2. Unlike the Clayton copula, which focuses on lower tail dependence, the Frank copula captures a symmetric dependence structure across the entire distribution. This means that the risks exhibit equal levels of dependence at both ends of the distribution, not concentrated in the tails.

The Frank copula is defined by the following generator function $\varphi(u)$:

$$\varphi(u) = -\frac{1}{\theta} \ln\left(\frac{1 + (e^{-\theta u} - 1)}{e^{-\theta} - 1}\right),$$

where θ is the copula parameter that governs the strength of the dependence.

The full Frank copula, which describes the joint survival function, is given by:

$$C_\theta(u_1, u_2) = \varphi^{-1}(\varphi(u_1) + \varphi(u_2))$$
$$= \frac{-\ln\left[1 + (e^{-\theta u_1} - 1)(e^{-\theta u_2} - 1)\right]}{e^{-\theta} - 1}, \tag{2}$$

where $u_1 = F_1(t_1)$ and $u_2 = F_2(t_2)$ are the uniform variables corresponding to the marginal cumulative distribution functions of T_1 and T_2, respectively. This formulation captures the dependence between the two risks, allowing for flexible modelling of their joint behaviour while maintaining symmetric dependence throughout the distribution.

4.3.2. *Case 2: Clayton*

The Clayton copula belongs to the Archimedean family and is well-suited for modelling lower tail dependence, a feature that allows for increased joint probability of early event times. This property is particularly relevant in financial contexts involving correlated prepayment and default risks. The copula is defined for $\theta > 0$ by the expression:

$$C_\theta(u_1, u_2) = \left(u_1^{-\theta} + u_2^{-\theta} - 1\right)^{-1/\theta}, \tag{3}$$

where $u_1 = F_1(t_1)$ and $u_2 = F_2(t_2)$ denote the marginal cumulative distribution values.

By combining the copula C_θ with specified marginals F_1 and F_2, a joint distribution over the latent failure times (T_1, T_2) is constructed. This allows the marginal behaviour of each risk to be modelled independently while incorporating flexible dependence between them.

4.4. *Estimation of Parameters using Maximum Likelihood Estimation*

For estimation, we consider n independent observations of the form (t_i, I_i), where $t_i \in \mathbb{R}_+$ is the observed event time and $I_i \in \{1, 2\}$ is an indicator denoting the cause of failure. The joint distribution of the latent event times (T_1, T_2) is defined via a copula function $C_\theta(u_1, u_2)$ with marginal distributions F_1 and F_2, where $u_j = F_j(t)$, and densities $f_j(t) = \frac{d}{dt} F_j(t)$, for $j = 1, 2$.

Under this framework, the likelihood contribution for individual i is determined by which event occurs first:

$$L_i(\boldsymbol{\theta}) = \begin{cases} f_1(t_i) \cdot \Pr(T_2 > t_i \mid T_1 = t_i), & \text{if } I_i = 1, \\ f_2(t_i) \cdot \Pr(T_1 > t_i \mid T_2 = t_i), & \text{if } I_i = 2, \end{cases} \tag{4}$$

To compute the conditional survival probabilities in the likelihood, we apply the copula representation. By Sklar's theorem, any joint distribution with continuous marginals can be expressed using a copula. Specifically, let

$$U_1 = F_1(T_1), \quad U_2 = F_2(T_2),$$

so that $(U_1, U_2) \in [0,1]^2$ has joint distribution function $C_\theta(u_1, u_2)$. The conditional distribution function of U_2 given $U_1 = u_1$ is given by

$$\Pr(U_2 \leq u_2 \mid U_1 = u_1) = \frac{\partial}{\partial u_1} C_\theta(u_1, u_2), \tag{5}$$

since the marginal property $C_\theta(u_1, 1) = u_1$ implies that the denominator in the general expression simplifies to unity.

Consequently, the conditional survival function becomes

$$\Pr(U_2 > u_2 \mid U_1 = u_1) = 1 - \frac{\partial}{\partial u_1} C_\theta(u_1, u_2). \tag{6}$$

Applying this to our setting, the conditional survival probabilities in the likelihood are expressed as:

$$\Pr(T_2 > t_i \mid T_1 = t_i) = 1 - \frac{\partial C_\theta}{\partial u_1} \left(F_1(t_i), F_2(t_i)\right), \tag{7}$$

$$\Pr(T_1 > t_i \mid T_2 = t_i) = 1 - \frac{\partial C_\theta}{\partial u_2} \left(F_1(t_i), F_2(t_i)\right) \tag{8}$$

4.4.1. *Case 1: Frank Copula*

Let $\boldsymbol{\theta} = (\theta, \alpha_1, \alpha_2, \beta_1, \beta_2)$ denote the vector of model parameters, where θ denotes the Frank copula dependence parameter, $\alpha_1, \alpha_2 > 0$ are the shape parameters of the Gamma distributions associated with loan prepayment and loan default respectively, and $\beta_1, \beta_2 > 0$ are the rate parameters for the Gamma distributions associated with loan prepayment and loan default risks, respectively.

The probability density functions for the Gamma distributions are:

$$f_j(t) = \frac{1}{\Gamma(\alpha_j)} t^{\alpha_j - 1} e^{-\frac{t}{\beta_j}} \beta_j^{\alpha_j}$$

for $j = 1, 2$.

The likelihood contribution for an individual observation (t_i, I_i) uses the general competing risks framework introduced in equation (4), with the conditional survival term derived from the properties of the frank copula.

The conditional survival probability (6) uses the Frank copula, so $C_\theta(u_1, u_2)$ has the same form as in equation (2). The partial derivative of $C_\theta(u_1, u_2)$ with respect to u_1 is:

$$\frac{\partial C_\theta(u_1, u_2)}{\partial u_1} = \frac{\theta e^{-\theta u_1}(e^{-\theta u_2} - 1)}{(e^{-\theta} - 1)\left[1 + \frac{(e^{-\theta u_1} - 1)(e^{-\theta u_2} - 1)}{e^{-\theta} - 1}\right]} \tag{9}$$

Substituting equation (9) into equation (6) gives:

$$\Pr(T_2 > t_i \mid T_1 = t_i) = 1 - \frac{\theta e^{-\theta u_1}(e^{-\theta u_2} - 1)}{(e^{-\theta} - 1)\left[1 + \frac{(e^{-\theta u_1} - 1)(e^{-\theta u_2} - 1)}{e^{-\theta} - 1}\right]} \tag{10}$$

with $i = 1, 2$.

Thus, the likelihood contribution for the i-th observation is:

$$L_i(\boldsymbol{\theta}) = \begin{cases} f_1(t_i) \cdot \left[1 - \frac{\partial C_\theta(u_1, u_2)}{\partial u_1}\right], & \text{if } I_i = 1, \\ f_2(t_i) \cdot \left[1 - \frac{\partial C_\theta(u_1, u_2)}{\partial u_2}\right], & \text{if } I_i = 2. \end{cases}$$

The full log-likelihood function is then:

$$\ln \mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln L_i(\boldsymbol{\theta}).$$

This log-likelihood is evaluated numerically and maximized subject to the constraints $\theta,\ \alpha_1,\ \alpha_2,\ \beta_1,\ \beta_2 > 0$

4.4.2. *Case 2: Clayton Copula*

In this specification, the dependence between T_1 and T_2 is modeled using the Clayton copula. Let the parameter vector be $\boldsymbol{\theta} = (\theta, \lambda_1, \lambda_2)$, where $\theta > 0$ denotes the copula dependence parameter, and $\lambda_1, \lambda_2 > 0$ are the constant hazard rates associated with loan prepayment and loan default, respectively. The marginal distributions are assumed to follow exponential laws with densities $f_j(t) = \lambda_j e^{-\lambda_j t}$, for $j = 1, 2$.

The likelihood contribution retains the structure in Equation (4), with the conditional survival term derived from the properties of the Clayton copula. For the Clayton copula, the bivariate copula function is defined as in Equation (3). Differentiating with respect to u_1, the conditional density simplifies to

$$\frac{\partial}{\partial u_1} C_\theta(u_1, u_2) = \left(u_1^{-\theta} + u_2^{-\theta} - 1\right)^{-(1+1/\theta)} u_1^{-\theta-1}. \tag{11}$$

Substituting Equation (11) into Equation (6) yields the conditional survival probability:

$$\Pr(U_2 > u_2 \mid U_1 = u_1) = 1 - \left(u_1^{-\theta} + u_2^{-\theta} - 1\right)^{-(1+1/\theta)} u_1^{-\theta-1}. \tag{12}$$

To express this in terms of the survival times T_1 and T_2, define the marginal distribution functions as

$$u_j = F_j(t) = 1 - e^{-\lambda_j t}, \quad j = 1, 2.$$

Substituting into Equation (12) yields the conditional survival probabilities required for the likelihood:

$$\Pr(T_2 > t \mid T_1 = t) = 1 - \left(u_1^{-\theta} + u_2^{-\theta} - 1\right)^{-(1+1/\theta)} u_1^{-\theta-1},$$

$$\Pr(T_1 > t \mid T_2 = t) = 1 - \left(u_1^{-\theta} + u_2^{-\theta} - 1\right)^{-(1+1/\theta)} u_2^{-\theta-1}.$$

Accordingly, for a sample of n independent observations $\{(t_i, I_i)\}_{i=1}^{n}$, the individual likelihood contributions are:

$$L_i(\boldsymbol{\theta}) = \begin{cases} \lambda_1 e^{-\lambda_1 t_i} \cdot \left[1 - \tilde{u} \cdot u_1^{-(1+\theta)}\right], & \text{if } I_i = 1, \\ \lambda_2 e^{-\lambda_2 t_i} \cdot \left[1 - \tilde{u} \cdot u_2^{-(1+\theta)}\right], & \text{if } I_i = 2, \end{cases}$$

where $\tilde{u}$ is given by:

$$\tilde{u} = \left(u_1^{-\theta} + u_2^{-\theta} - 1\right)^{-(1+\frac{1}{\theta})}$$

The full log-likelihood function is then:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log L_i(\boldsymbol{\theta}),$$

This log-likelihood is evaluated numerically and maximized subject to the constraints $\theta > 0$, $\lambda_1 > 0$, and $\lambda_2 > 0$.

4.5. Nonparametric Bootstrap for Parameter Uncertainty

While Maximum Likelihood Estimators provide point estimates for parameters such as θ, λ, and α, they do not yield confidence intervals directly. To quantify uncertainty in the maximum likelihood estimates $\widehat{\boldsymbol{\theta}}$, we apply a nonparametric bootstrap procedure. Let:

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ \ell(\boldsymbol{\theta})$$

For each bootstrap iteration $b = 1, \dots, B$:

1. A bootstrap sample $\{(t_i^{*b}, I_i^{*b})\}_{i=1}^n$ is drawn with replacement from the original data;
2. The log-likelihood $\ell^{*b}(\theta)$ is computed on the resampled data;
3. The corresponding estimate is:

$$\widehat{\boldsymbol{\theta}}^{*b} = \arg\max_{\boldsymbol{\theta}} \ \ell^{*b}(\boldsymbol{\theta})$$

This results in a bootstrap distribution $\{\widehat{\boldsymbol{\theta}}^{*1}, \dots, \widehat{\boldsymbol{\theta}}^{*B}\}$. For each parameter θ_j, a percentile-based $(1-\alpha)$ confidence interval is obtained via:

$$\left[\widehat{\theta}_j^{*,\alpha/2}, \ \widehat{\theta}_j^{*,1-\alpha/2}\right]$$

where $\widehat{\theta}_j^{*,q}$ denotes the empirical q-th quantile of the bootstrap estimates $\{\widehat{\theta}_j^{*b}\}_{b=1}^B$.

This approach facilitates model-independent quantification of parameter uncertainty and supports direct comparison of estimation variability across alternative copula formulations.

5. Simulation Study

5.1. Part I: Frank Copula

5.1.1. Data Generation

We examine the use of a Frank copula to model dependency between two competing risks: loan default (the borrower failing to pay) and loan repayment (the borrower successfully repaying the loan in full). The Frank copula provides a symmetric dependence structure, meaning that dependence is modelled evenly across the entire distribution — neither early nor late loan closures are disproportionately influential.

This setting reflects real-world scenarios where the two loan closure types may be influenced by shared economic or market factors, without a strong directional bias. For example, improving economic conditions may simultaneously reduce default risk and accelerate repayment, or worsening conditions may delay repayment and increase default likelihood, meaning that both outcomes are similarly impacted across time. The Frank copula is ideal in such cases as it lacks tail dependence and instead models a more balanced, global dependency between variables.

The two loan times are drawn from gamma distributions, where both loan default and repayment are assigned the same marginal parameters to easily see patterns caused by their dependency: shape = 2 and rate = 4, corresponding to a mean closure time of 0.5 for each component.

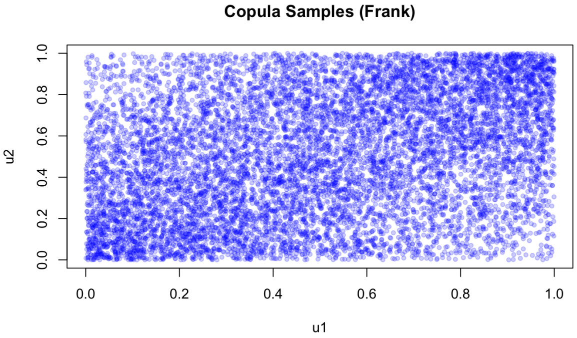


Figure 2. Scatterplot of Frank copula samples (u_1, u_2)

A scatterplot of the sampled copula pairs (u_1, u_2) (shown above) confirms the symmetric nature of the Frank copula. There is no strong clustering in the tails — instead, points are evenly dispersed around the diagonal, suggesting uniform correlation across the support. This aligns with the Frank copula's property of no tail dependence, but still permits moderate association between components.

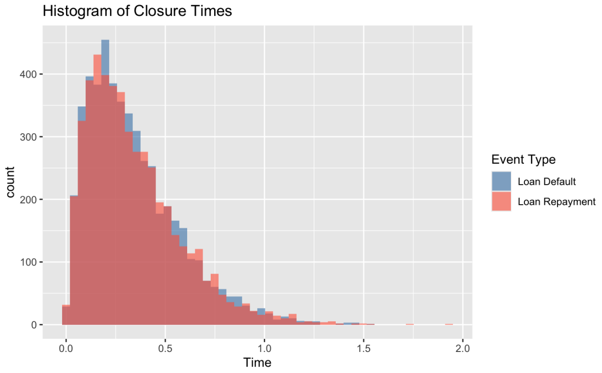


Figure 3. Histogram of closure times by event type under Frank copula.

A histogram of closure times coloured by event type shows the distribution of observed loan closures from loan default and loan repayment. Because both marginals are identically distributed, the differentiation between causes is driven entirely by the dependency structure introduced by the copula, where the distribution of causes is more balanced over time.

5.1.2. Fit Marginals

Fitting the marginals for each loan closure cause using a selection of candidate distributions (exponential, weibull, gamma, and log-normal), the gamma distribution was correctly identified as the best fit in both cases. This allows us to move forward with parameter estimation under correct distributional assumptions.

These plots help visually assess how well the fitted distributions match the empirical data.

Q-Q plots comparing the empirical data to the fitted gamma marginals support this result visually. Both loan default and loan repayment risks follow the gamma quantiles closely, with minor deviations in the upper tail.

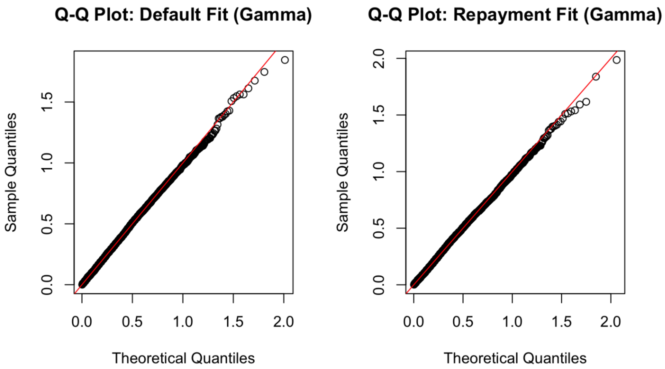


Figure 4. Q-Q plots comparing empirical data to gamma marginals for both loan default and prepayment.

5.1.3. Estimate Parameters

The Frank copula was chosen for this stage of the model because it captures symmetric dependence between risks, meaning that association between loan closure times is modelled consistently across the entire distribution, rather than being concentrated in the tails. This makes it particularly appropriate for competing risks scenarios where the interaction between loan closure types is not limited to early or extreme events.

We estimate the copula dependence parameter (θ), shape parameters (α_1, α_2) and rate parameters (β_1, β_2) for the Gamma distributions using maximum likelihood estimation on the full synthetic data set.

To assess uncertainty in the parameter estimates, we use bootstrap resampling. This gives a distribution of estimates for $\theta, \alpha_1, \beta_1, \alpha_2$, and β_2.

We now visualise the distribution of bootstrap estimates for each parameter. Each plot includes a black dashed line for the known true value used to generate the synthetic data, and a red solid line representing the MLE point estimate from the full dataset. A legend is included for clarity.

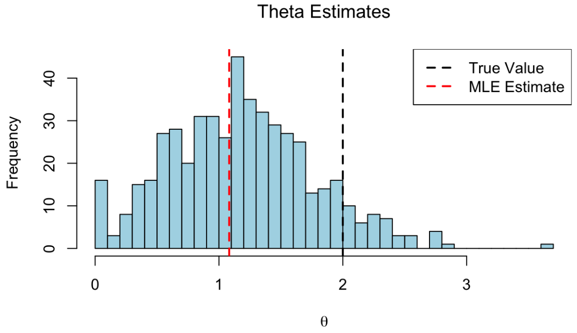
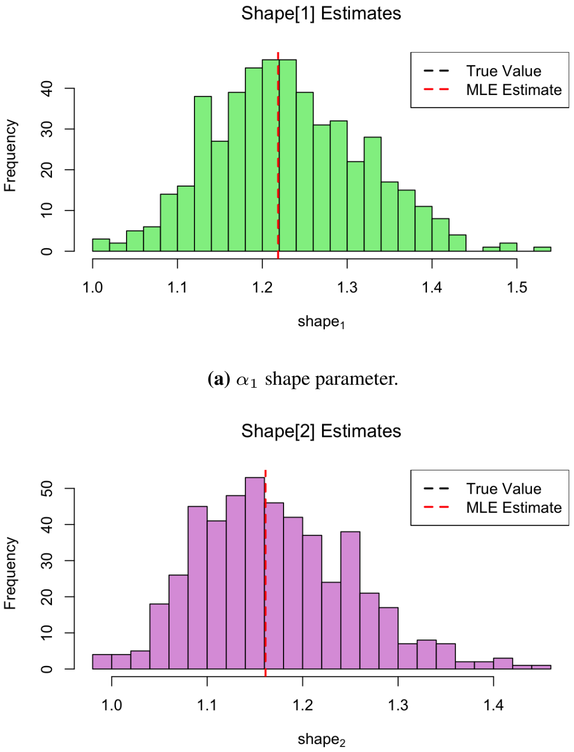


Figure 5. Bootstrap distribution of the Frank copula dependence parameter θ.



(a) α_1 shape parameter.

(b) α_2 shape parameter.

Figure 6. Bootstrap distribution of the Gamma shape parameters for loan prepayment and loan default respectively.

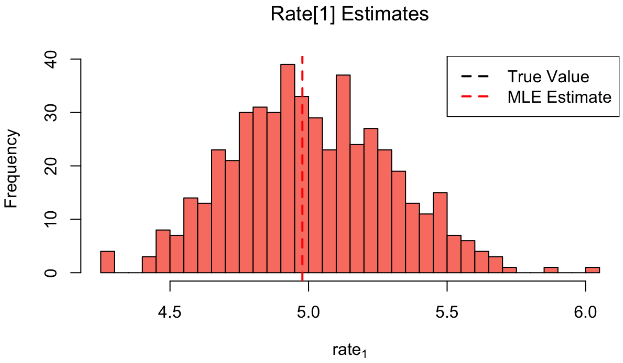


Figure 7. Bootstrap distribution of the Gamma rate parameter β_1 for loan repayment.

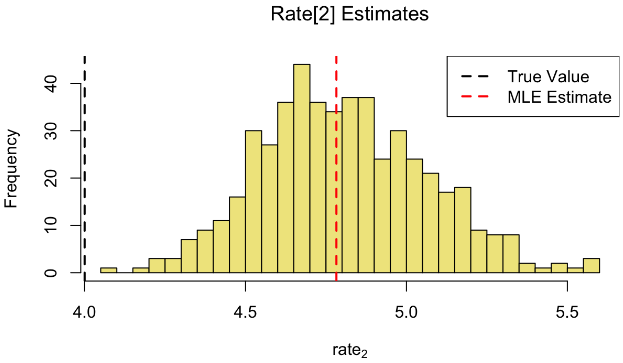


Figure 8. Bootstrap distribution of the Gamma rate parameter β_2 for loan default.

5.1.4. Discussion

This synthetic implementation serves as a controlled environment to test and demonstrate the use of copulas — specifically the Frank copula — in modelling competing risks. By simulating data with known dependence and marginal characteristics, we were able to assess the accuracy of parameter estimation methods and validate their performance before applying them to real-world data.

We generated 10,000 observations representing time-to-event data for two competing risks: loan defaults and loan repayments. Dependence between the two loan closure types was introduced using a Frank copula with a known parameter of $\theta = 2$, chosen for its ability to model symmetric dependence — appropriate for scenarios where shared risk factors influence both causes consistently over time, without concentrating in the tails.

The marginal distributions for both risks were set to gamma, with $\alpha_1 = 2, \beta_1 = 4$ for loan default and $\alpha_2 = 2, \beta_2 = 4$ for loan repayment. Observed closure times were the minimum of the two, with the cause recorded as the corresponding event type.

Notably, from the above, we can make some key observations

- The Frank copula generated balanced dependence, visible as even dispersion around the diagonal in the u_1–u_2 scatterplot, without clustering in the tails.
- Loan closures from both causes were distributed relatively evenly over time, consistent with the identical gamma marginals.
- Gamma distribution assumptions for both causes were visually supported, with empirical quantiles aligning well with the fitted distributions.
- The distributions of bootstrap estimates for $\theta, \alpha_1, \beta_1, \alpha_2$ and β_2 were all centred near the MLE estimates with low spread. Visuals included the true parameter values and MLE estimates for reference.
- MLEs showed a consistent pattern of underestimating the true shape parameters and θ, with slight overestimation of the rate parameters. This reflects moderate bias, particularly in the copula component.

Table 1. Comparison of MLE, Bootstrap, and True Values

Parameter	Truth	MLE	Bootstrap Mean
θ	2.00	1.08	1.18
α_1	2.00	1.22	1.23
β_1	4.00	4.98	5.00
α_2	2.00	1.16	1.17
β_2	4.00	4.78	4.80

- Bootstrap means confirmed the stability of the MLE estimates, with distributions tightly clustered around the MLEs, but not the true values, due to the bias of our estimates.
- The underestimation of θ likely stems from the Frank copula's lack of tail dependence, which makes dependence harder to detect in joint extremes, especially with moderate true dependence and gamma marginals.

5.1.5. Conclusions

This synthetic implementation demonstrated that copula-based competing risks models can reasonably recover marginal and dependence parameters when using the Frank copula, though with increased estimation uncertainty. The results support the viability of using Frank copulas for modelling symmetric dependence, particularly in scenarios without strong tail effects.

What went well:

- The estimation process was robust in both MLE and bootstrapping.
- Visual checks supported the use of gamma marginals.

Limitations:

- Estimates for θ and the gamma shape parameters showed both bias and greater spread, likely due to the Frank copula's lack of tail dependence, highlighting limitations of full likelihood estimation in this setup.
- The symmetric dependence structure of the Frank copula may make identification harder when dependence is moderate and not concentrated in the tails.
- Shape parameters for both marginals were also underestimated, while rate parameters were overestimated, suggesting a parameter trade-off: the model may have compensated for lighter tails (lower shape) by increasing the rate to maintain similar means.
- There is increased estimation complexity. Five parameters are being estimated simultaneously. The gamma distribution involves two interacting parameters (shape and rate), which can lead to identifiability issues. Small changes in one can compensate for changes in the other, making precise estimation harder.

Despite these limitations, the simulation provides valuable insight into the strengths and challenges of using the Frank copula for competing risks modelling, and forms a useful benchmark for further exploration with real-world data or alternative copulas, perhaps with stronger tail dependency, such as the Clayton copula.

5.2. Part II: Clayton Copula & Exponential Marginals

5.2.1. Data Generation

We simulate two dependent failure times (T_1, T_2) representing loan repayment and loan default (the competing risks) using a Clayton copula to induce positive dependence.

A scatter plot of copula samples (u_1, u_2) illustrates the dependence structure imposed by the Clayton copula.

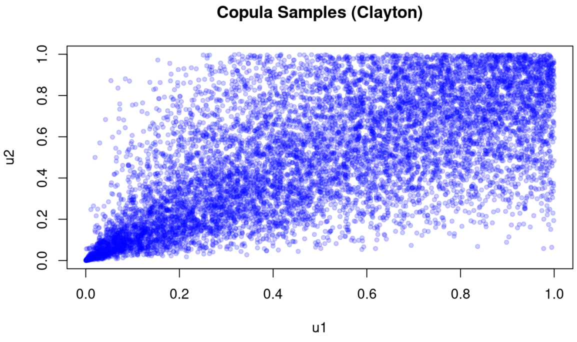


Figure 9. Scatterplot of Clayton copula samples (u_1, u_2).

A histogram of closure times colored by event type shows the distribution of observed closures from repayment and default.

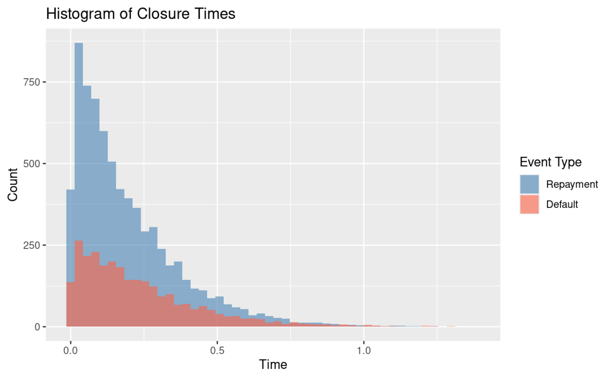


Figure 10. Histogram of closure times under Clayton copula.

5.2.2. *Fit Marginals*

We split the simulated data into two groups by event type and fit several candidate distributions to each. Model selection is based on AIC, and we use shape and AIC difference tolerances to determine if an exponential fit can acceptably replace a Weibull.

QQ plots help visually assess how well the fitted distributions match the empirical data.

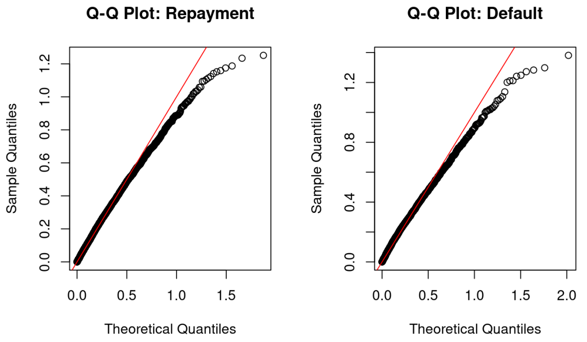


Figure 11. Q-Q plots comparing empirical data to fitted exponential marginals.

5.2.3. *Estimate Parameters*

The Clayton copula was chosen for the final model because it models lower tail dependence, which is appropriate for competing risks scenarios where extreme (early) failure in one component may increase the likelihood of failure in the other. This reflects real-world settings where risks may be correlated during stress events, making Clayton a natural and interpretable choice

We estimate the copula dependence parameter (θ) and exponential rates (λ_1, λ_2) using maximum likelihood estimation on the full synthetic dataset.

To assess uncertainty in the parameter estimates, we use bootstrap resampling. This gives a distribution of estimates for θ, λ_1, and λ_2.

We can now visualize the distribution of bootstrap estimates for each parameter. Each plot includes a black dashed line for the known true value used to generate the synthetic data, and a red solid line representing the MLE point estimate from the full dataset.

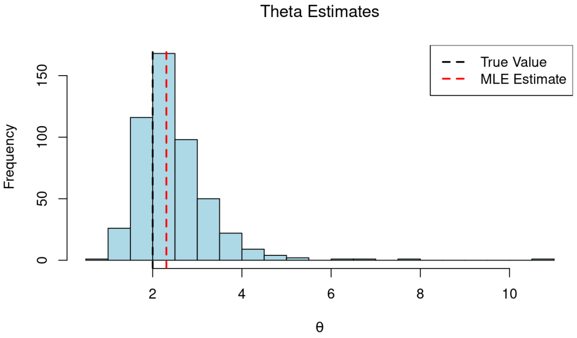
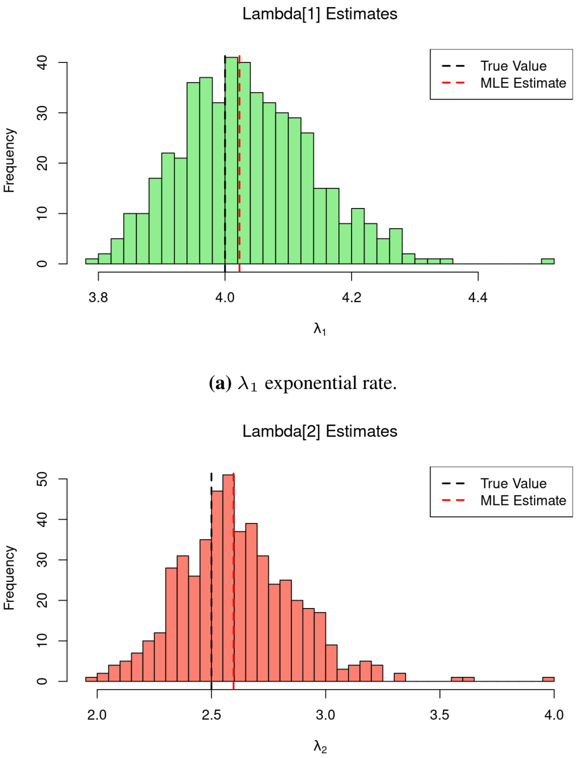


Figure 12. Bootstrap distribution of the Clayton copula parameter θ.



(a) λ_1 exponential rate.

(b) λ_2 exponential rate.

Figure 13. Bootstrap distribution of the exponential rates for loan repayment and loan default respectively.

5.2.4. *Discussion*

This synthetic implementation serves as a controlled environment to test and demonstrate the use of copulas — specifically the Clayton copula — in modelling competing risks. By simulating data with known dependence and marginal characteristics, we were able to assess the accuracy of parameter estimation methods and validate their performance.

We generated 10,000 observations representing time-to-event data for two competing risks: loan repayment and loan default. Dependence between the two failure types was introduced using a Clayton copula with a known parameter of $\theta = 2$, chosen for its ability to model lower tail dependence, which is appropriate for scenarios where early failure in one increases the risk of failure in another.

The marginal distributions for both risks were set to exponential, with $\lambda_1 = 4$ for repayment and $\lambda_1 = 2.5$ for default. Observed closure times were the minimum of the two, with the cause recorded as the corresponding event type.

From what was previously mentioned, some key observations can be obtained made:

- As evident from figure [reference], the Clayton copula generated clear lower tail dependence, visible as clustering in the bottom-left of the u_1–u_2 scatter plot.
- As can be seen in figure [reference], repayments dominated early in the timeline, while default events occurred more evenly across time, consistent with their lower hazard rate.
- From figure [ref], we can see that the exponential distribution assumptions for both causes were visually supported. The points closely followed the theoretical line, with only mild deviation in the upper tail, confirming good model fit.
- The distributions of bootstrap estimates for θ, λ_1, and λ_2 were all centred near their true values, with low spread.

Table 2. Comparison of MLE, Bootstrap, and True Values

Parameter	**Truth**	**MLE**	**Bootstrap Mean**
θ	2.00	2.31	2.45
λ_1	4.00	4.02	4.04
λ_2	2.50	2.60	2.62

- The MLEs were very close to the true values for all parameters, indicating successful recovery of model inputs and the bootstrap means confirmed the stability of these estimates, with distributions tightly clustered around both the true values and the MLEs.
- θ showed slightly more variability across bootstrap samples, which is expected due to its more complex influence on joint behaviour. Unlike the marginal rate parameters, which depend only on the individual time distributions, θ governs the dependence structure between the risks, making it more sensitive to variation in the joint tails of the data. This results in greater estimation uncertainty, especially when dependence is moderate or when sample fluctuations disproportionately affect joint failures.

This synthetic implementation successfully demonstrated that copula-based competing risks models can recover true marginal and dependence parameters with high accuracy, under ideal conditions. The results validated both the model structure and estimation approach, particularly the use of MLE and bootstrap for parameter inference.

With that being said, θ exhibited more spread in bootstrap results, suggesting that dependence parameters may be less stable under resampling. Real-world data may also introduce noise, censoring, or model violations not present in this controlled setup.

Nonetheless, this simulation provides a strong foundation for applying the same framework to empirical data.

6. Case Study

To truly ascertain the suitability of copulas in modelling the relationship between two risk factors, namely the competing risks of defaulting on a loan and prepayment, a real world dataset is studied in this section.

To avoid adding further uncertainty in our estimates via imputation or other similar methods, a financial data set was sought that included origination and termination dates, the maturity date, alongside a clear cause of termination. One such suitable dataset was found that included these restrictions with minimal missing-ness [14]. Namely, the *Lending Club loan data* dataset includes all of the $2,260,701$ loans approved on the US based peer to peer lending site *Lending Club* from the beginning of 2007 to 2018.

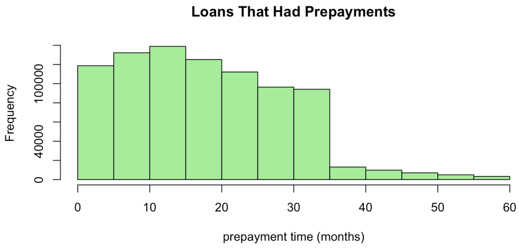
The dataset contains three loan status indicators of particular interest - 'Fully Paid' indicating that the loan was fully settled up in the study period, 'Default' when the loan has been in arrears for more than 120 days, and 'Charged Off' when the loan has been officially deemed written off as a bad debt. For this exercise, loans with status indicators of either 'Default' or 'Charged Off' will henceforth be referred to as defaulted. For loans that were described as 'Fully Paid', a further indicator was derived. Chiefly, if the maturity period of a loan was longer than the time it took the lendee to repay it, then such an observation is defined as a prepayment.

6.1. *Estimating Marginals*

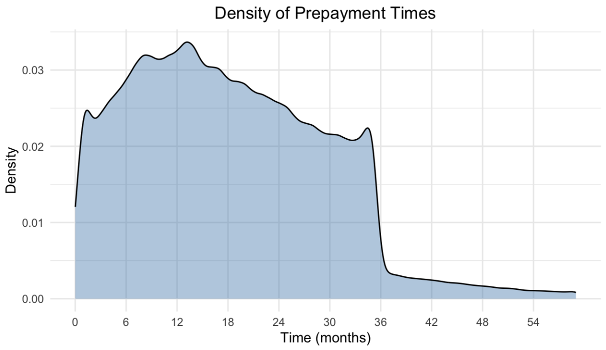
After minimal data cleaning, the dataset was temporarily split into two categories - observations where the loan ended due to prepayment, and those that ended due to default. The 'arrival time' of these occurrences were calculated by finding the difference (in months) of the start date of the loan and the date where the data was censored. Then, these two groups of arrival times were used to estimate their respective marginal distributions, as described in the following section.

6.1.1. *Loan Prepayment Marginal Estimation*

The histogram of the prepayment arrival times is:



and their density was graphed using R:



The histogram and densities describe a roughly unimodal distribution with a long right tail distribution. Notably, the distribution is not monotonically decreasing on the range of values for time - indicating that an exponential marginal can be ruled out for this specific real world dataset. Hence, a Weibull, Gamma, and Log-Normal distribution are fitted to the time observations using Maximum Likelihood Estimation methods as their respective characteristics could plausibly explain the data. Out of the three contenders, the Weibull distribution estimation had the lowest AIC score with parameters:

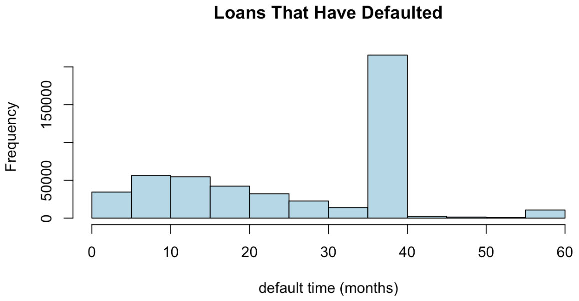
Hence, the loan prepayment time (T_1) marginal distribution f_1 is assumed to be Weibull distributed with shape parameter 1.512 and scale parameter 19.903. This marginal distribution will contribute to the competing risks analysis.

Parameter	Estimate (95 % CI)
shape (α)	1.512 [1.509–1.515]
scale (λ)	19.903 [19.875–19.931]

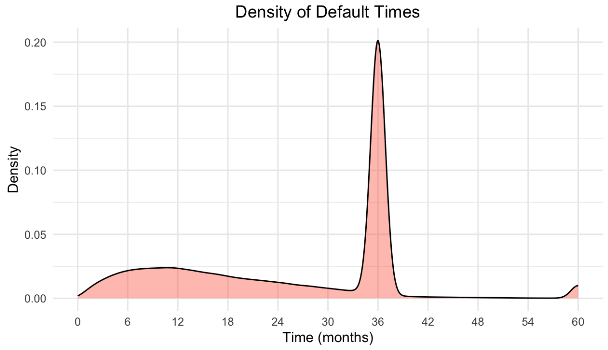
Table 3. Weibull parameter estimates and 95 % confidence intervals for prepayment times.

6.1.2. *Loan Default Marginal Estimation*

Next, the subset of times where the indicator was that the loan was prepayed is considered. They were plotted on a histogram as follows:



The occurrence of defaults seems to be more clustered than the occurrence of prepayments in this dataset. The density of the default times (T_2) confirms this belief:



There are clearly three local peaks in the density around the time of the one, three, and 5 year mark. Given there is more than one distinct peak in this density, it could be described as being a multimodal distribution. The observations are fitted by single Weibull, Gamma, and Log-Normal distributions. Again, the AIC for the Weibull distribution has the lowest value for the default time observations. That said, mixtures of the these distributions, alongside the exponential distribution, for various component sizes were also considered. The AIC method penalises the likelihood of the fit against the amount of parameters being estimated in the model. Even after this penalisation, the marginal distribution with a mixture of three Weibulls is found to have the best fit with tight confidence intervals: Meaning the most explanatory p.d.f. of the marginal distribution of the n default times is:

Component	Weight (π)	Shape (α)	Scale (λ)
1	0.1113292	1.533138	29.57184
2	0.4072534	1.720762	16.63987
3	0.4814174	1353.704117	36.00209

Table 4. Three-component Weibull mixture parameter estimates.

$$f_2 = \mathbb{P}(t_n) = \sum_{k=1}^{3} \pi_k \cdot \mathbb{P}_{\text{Weibull}}(t_n \mid \alpha_k, \lambda_k)$$

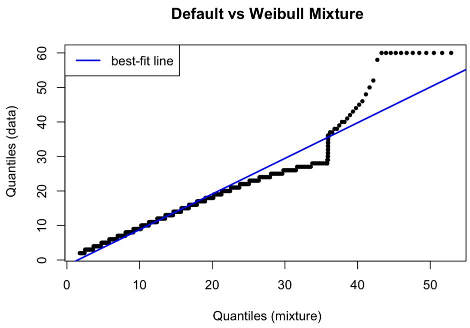


Figure 14. Q-Q plot of Mixture of 3 Weibull Distributions versus the Data's Distribution

where $\mathbb{P}_{\text{Weibull}}$ is the density of the Weibull distribution. The Q-Q plot showed that the 3 mixture made a good attempt at capturing most of the distribution - but the large spike around 36 months is clearly hard to model without over-fitting the model.

6.2. *Conditional survival probabilities under the Frank, Clayton and Gumbel copulas*

Now, the relationship between the two competing risks of loan default and loan prepayment within our dataset will be investigated by fitting Gumbel, Frank, and Clayton copulas - assuming the marginal densities f_1 and f_2 derived in the previous subsection. These copulas in particular were chosen given the evidence Given the data is censored, pseudo observations are constructed using the marginal's survival functions:

$$F_1(t_n) = u_1 = 1 - S_1(t_n)$$
$$F_2(t_n) = u_2 = 1 - S_2(t_n)$$

and as an extension of Sklar's Theorem, we see from Section 4.2 that:

$$\mathbb{P}(T_2 > t_n \mid T_1 = t_n) = 1 - \frac{\partial}{\partial u_1} C\left(F_1(u_n), F_2(u_n)\right)$$
$$\mathbb{P}(T_1 > t_n \mid T_2 = t_n) = 1 - \frac{\partial}{\partial u_2} C\left(F_1(u_n), F_2(u_n)\right)$$

Hence, the individual log likelihood contribution for the n'th observation with indicator variable I_n is:

$$L_n(\boldsymbol{\theta}) = \begin{cases} f_1(t_n) \cdot \Pr(T_2 > t_n \mid T_1 = t_n), & \text{if } I_n = 1, \\ f_2(t_n) \cdot \Pr(T_1 > t_n \mid T_2 = t_n), & \text{if } I_n = 2, \end{cases}$$

The Frank and Clayton copula's dependence functions for two variables u_1 & u_2 were fully derived in Section 4.3, and the Gumbel Copula dependence needs to be derived in this section:

Frank Copula Using equation (9) and equation (6), we can derive the survival time estimation for Frank Copula:

$$\Pr(T_2 > t_i \mid T_1 = t_i) = 1 - \frac{\theta e^{-\theta u_1}(e^{-\theta u_2} - 1)}{(e^{-\theta} - 1)\left[1 + \frac{(e^{-\theta u_1} - 1)(e^{-\theta u_2} - 1)}{e^{-\theta} - 1}\right]} \tag{13}$$

with $i = 1, 2$.

Thus, the likelihood contribution for the i-th observation is:

$$L_i(\boldsymbol{\theta}) = \begin{cases} f_1(t_i) \cdot \left[1 - \frac{\partial C_\theta(u_1, u_2)}{\partial u_1}\right], & \text{if } I_i = 1, \\ f_2(t_i) \cdot \left[1 - \frac{\partial C_\theta(u_1, u_2)}{\partial u_2}\right], & \text{if } I_i = 2. \end{cases}$$

The values from this piecewise function are then added to the full log-likelihood function:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log L_i(\boldsymbol{\theta}),$$

Clayton Copula: yields the conditional survival probabilities required for the likelihood:

$$\Pr(T_2 > t \mid T_1 = t) = 1 - \left(u_1^{-\theta} + u_2^{-\theta} - 1\right)^{-(1+1/\theta)} u_1^{-\theta-1},$$
$$\Pr(T_1 > t \mid T_2 = t) = 1 - \left(u_1^{-\theta} + u_2^{-\theta} - 1\right)^{-(1+1/\theta)} u_2^{-\theta-1}.$$

Accordingly, for a sample of n independent observations $\{(t_i, I_i)\}_{i=1}^{n}$, the individual likelihood contributions are:

$$L_i(\boldsymbol{\theta}) = \begin{cases} f_1(t_i) \cdot \left[1 - \tilde{u} \cdot u_1^{-(1+\theta)}\right], & \text{if } I_i = 1, \\ f_2(t_i) \cdot \left[1 - \tilde{u} \cdot u_2^{-(1+\theta)}\right], & \text{if } I_i = 2, \end{cases}$$

where $\tilde{u}$ is given by:

$$\tilde{u} = \left(u_1^{-\theta} + u_2^{-\theta} - 1\right)^{-(1+\frac{1}{\theta})}$$

with f_1 and f_2 are the marginal densities of the arrival time variables.

Gumbel Copula: The Gumbal Copula has the generator function $\varphi(u) = (-\log(u))^{\theta}$, which implies the inverse generator function is $\varphi^{-1}(u) = \exp(-u^{1/\theta})$. Hence, the full Frank copula $C_\theta(u_1, u_2)$, which quantifies the joint survival function of the two psuedo observations, is given by:

$$\begin{aligned} C_\theta(u_1, u_2) &= \varphi^{-1}(\varphi(u_1) + \varphi(u_2)) \\ &= \exp\left[-\left((-\log(u_1))^{\theta} + (-\log(u_2))^{\theta}\right)^{1/\theta}\right] \end{aligned}$$

And the partial derivatives of the Copula with respect to the pseudo observations is

$$\frac{\partial C}{\partial u_1} = C_\theta(u_1, u_2)\, \frac{(-\ln u_1)^{\theta-1}}{u_1}\, \left[(-\ln u_1)^{\theta} + (-\ln u_2)^{\theta}\right]^{\frac{1}{\theta}-1}$$
$$\frac{\partial C}{\partial u_2} = C_\theta(u_1, u_2)\, \frac{(-\ln u_2)^{\theta-1}}{u_2}\, \left[(-\ln u_1)^{\theta} + (-\ln u_2)^{\theta}\right]^{\frac{1}{\theta}-1}$$

which allows us to define the conditional survival function as:

$$\begin{aligned} &\Pr(U_k > u_k \mid U_j = u_j) \\ &\qquad = 1 - \frac{\partial C_\theta(u_k, u_j)}{\partial u_j} \\ &\qquad = C_\theta(u_k, u_j)(-\ln u_k)^{\theta-1} \cdot \frac{\left[(-\ln u_k)^{\theta} + (-\ln u_j)^{\theta}\right]^{\frac{1}{\theta}-1}}{u_k} \end{aligned}$$

Now that the various closed-form expressions for the conditional survival probabilities under the Frank, Clayton and Gumbel copulas, our competing-risks likelihood framework is fully devised. In the next section we will plug these into a numerical optimiser to jointly estimate the copula dependence parameter θ and the marginal distribution parameters.

7. Parameter Estimation

7.1. Introduction

Although all three copulas were fitted to the prepayment and arrival times, the copula that returned the best AIC score i.e. it explained the data the best of of these three choices, was the Frank copula. This is presumably due to the spike defaults at the time 36 months. Qualitatively, based on the histograms, there seems to be a strong dependence between the two competing risks. If you 'survive' pre-paying up until roughly the three year point, then you're then quite likely to default there. Given the Frank copula is the only one out of all of them that allows for negative dependence, it is plausible that it was the best fitting. The Clayton copula scored moderately well, given the apparent positive relationship between the two competing risks in over the first tail. That said, the magnitude of the change in correlation between the two competing risks in the upper tail was left unexplained by the Clayton copula - a penalisation in comparison to the relationship described by the Frank copula, which strictly models codependency only. This is advantageous for our data set, as the pseudo-observations (once constructed) are fairly diffuse with no strong tail dependence visible.

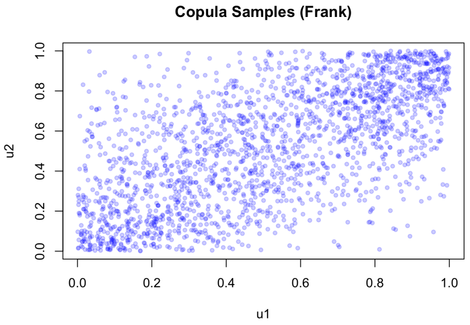


Figure 15. Scatter-plot of the fitted Frank copula pseudo-observations (u_1, u_2)

7.2. Results

To fit this copula to our loan data set, both of the marginal distributions and the Frank copula parameter θ. When estimated alongside the copula, the structure of the marginals changed (i.e. the marginal for defaults remained as a mixture of three Weibull distributed components), but the estimation of the parameters slightly changed. For the Clayton and Gumbel copula tail dependency parameters, their bootstrapped estimates had confidence intervals that were barely statistically different from zero, albeit the estimated dependency parameter for the Clayton copula was valued higher. Next, the Gumbel copula was fitted to the loan data. Through Maximum Likelihood Estimation, the best value for the Gumbel copula was found. After this, it was bootstrapped 500 times to attempt to get a clearer picture of its distribution and to ascertain its value in the estimation of the codependence between the pseudo observations. The parameter θ was estimated at 1.78 with a 95% confidence interval of [1.55, 2.05]. The bootstrapped estimates for the weight, shape and scale parameters of the Weibull distributions were well within the a well grounded and well fitting of the data.

	Copula Parameter Estimation	**95% CI**
Gumbel	**1.052**	**[1.0,1.11]**
Clayton	**0.476**	**[0.301,0.553]**

Table 5. Copula parameter estimates for the Gumbel and Clayton copulas and their 95 % confidence intervals.

7.3. Conclusions

The results of this section prove that Copulas are indeed a valid and flexible tool to use in risk and credit modelling. Notably, even with the multimodal distribution of one of the arrival times, the model had a decent fit. As a result of this model, we can say with confidence that a borrower who had defaulted by the median time in their loan had a 7 percentage points higher chance of having more likely to have pre-paid by that same time than an average borrower:

$$\Pr(u_1 \leq 0.5 \mid u_2 \leq 0.5) = \frac{C(0.5,0.5)}{0.5} \approx 1.07$$

This confirms the qualitative theory explained at the beginning of this section that prepayment and default seem to go hand in hand in this dataset.

8. Conclusion

This work set out to examine the use of copulas in modelling the dependence structure between competing financial risks. Beginning with a theoretical foundation, this paper introduced copulas as flexible tools for capturing dependency beyond what traditional multivariate distributions would allow. We then outlined several families of copulas and applied them within a competing risks framework. Using simulated data, the project demonstrated how different copulas could model symmetric and lower tail dependence, respectively, with parameter estimation carried out via maximum likelihood and uncertainty assessed through bootstrap methods. Finally, this theoretical framework and background was applied to a real-world dataset from Lending Club, with suitable marginal distributions estimated for each risk type.

In particular, the simulation studies detailed in section 5, provided a controlled environment to evaluate the effectiveness of copula-based models in capturing the dependence between competing risks. The Frank copula simulation, using gamma marginals, highlighted that symmetric dependence can be modelled reasonably well, though with some bias in the estimation of the dependence parameter θ, particularly due to the lack of tail dependence. In contrast, the Clayton copula simulation with exponential marginals showed strong performance in modelling lower tail dependence, recovering true parameter values with high accuracy and low variance. Across both setups, bootstrap resampling confirmed the stability of maximum likelihood estimates, though it also highlighted greater uncertainty in the dependence parameter compared to marginal parameters. These simulations validate that copulas, when appropriately chosen, can effectively capture complex dependency structures in competing risks settings.

The case study in section 6, real-world loan data was used to evaluate the suitability of copula-based methods. The most notable finding was that the distribution of default times was multimodal, indicating that simple parametric forms such as the exponential or single Weibull were inadequate. Accordingly, the marginal distributions were best estimated using a Weibull distribution for prepayments and a three-component Weibull mixture for defaults. This mixture model offered a much better fit and did not appear to introduce over-fitting. Three copula families, Gumbel, Frank, and Clayton, were then fitted to the data to model dependence, with the aim of identifying which structure best captured the joint behaviour of the risks. In general, the case study reinforced the value of flexible copula models and data-driven marginal selection in real-world financial risk settings, particularly where complex or multimodal event time distributions are observed.

Overall, the results from this paper support the use of copula-based methods in financial risk analysis, especially when traditional assumptions of independence or simple parametric forms are violated by empirical data. Future work could explore the integration of censoring and covariate effects, as well as the use of tail-sensitive copulas in more volatile financial environments. It could also focus on either mixture modelling or Hidden Markov modelling to attempt to better quantify the dynamic nature of the relationship between two competing risks over time.

9. GitHub

The code used to conduct the simulation studies and the empirical case study involving Lending Club presented in this paper is available in a public GitHub repository for transparency and reproducibility. All R scripts used for data generation, copula fitting, visualisation, and statistical evaluation can be accessed at https://github.com/conorcaseyc/copulas-competing-risks.

References

[1] D. Acemoglu, A. Ozdaglar, and A. Tahbaz-Salehi. "Systemic risk and stability in financial networks". In: *American Economic Review* 105.2 (2015), pp. 564–608.

[2] P. C. Austin, D. S. Lee, and J. P. Fine. "Introduction to the analysis of survival data in the presence of competing risks". In: *Circulation* 133.6 (2016), pp. 601–609.

[3] Delson Chikobvu and Owen Jakata. "Quantifying Diversification Effects of A Portfolio Using the Generalised Extreme Value Distribution- Archimedean Gumbel Copula Model". In: *Applied Mathematics & Information Sciences* 17.6 (Nov. 1, 2023), pp. 967–981. ISSN: 19350090, 23250399. DOI: 10.18576/amis/170603. URL: https://www.naturalspublishing.com/Article.asp?ArtcID=27706 (visited on 05/09/2025).

[4] FasterCapital. *Clayton Copula: A Journey Through Dependency Landscapes*. Accessed: 2025-05-09. Apr. 2025. URL: https://fastercapital.com/content/Clayton-Copula--The-Clayton-Copula--A-Journey-Through-Dependency-Landscapes.html.

[5] J. Jagtiani and W. W. Lang. "Strategic default on first and second lien mortgages during the financial crisis". In: *Journal of Fixed Income* 20.4 (2011), p. 7.

[6] M. Jordan. *Introduction to Copulas*. https://www.youtube.com/watch?v=R_7Qvbrb0jE&t=6s. YouTube video. Apr. 2020.

[7] Lie-Jane Kao, Po-Cheng Wu, and Cheng Few Lee. "An Assessment of Copula Functions Approach in Conjunction with Factor Model in Portfolio Credit Risk Management". In: *Handbook of Investment Analysis, Portfolio Management, and Financial Derivatives*. Chap. Chapter 16, pp. 573–591. DOI: 10.1142/9789811269943_0016. eprint: https://www.worldscientific.com/doi/pdf/10.1142/9789811269943_0016. URL: https://www.worldscientific.com/doi/abs/10.1142/9789811269943_0016.

[8] A. Meucci. "A Short, Comprehensive, Practical Guide to Copulas". In: *GARP Risk Professional* (2011), pp. 22–27.

[9] R. B. Nelsen. *An Introduction to Copulas*. Springer, 1999.

[10] E. B. Ntiamoah et al. "Loan default rate and its impact on profitability in financial institutions". In: *Research Journal of Finance and Accounting* 5.14 (2014), pp. 67–72.

[11] Tino Productions. *A Simple Introduction to Copulas*. https://www.youtube.com/watch?v=WFEzkoK7tsE. YouTube video. Mar. 2021.

[12] F. Salmon. *Recipe for Disaster: The Formula that Killed Wall Street*. https://www.wired.com/2009/02/wp-quant/. Feb. 2009.

[13] Pravin K Trivedi, David M Zimmer, et al. "Copula modeling: an introduction for practitioners". In: *Foundations and Trends® in Econometrics* 1.1 (2007), pp. 1–111.

[14] wordsforthewise. *All Lending Club loan data: 2007 through current Lending Club accepted and rejected loan data*. Kaggle Dataset. Retrieved May 7, 2025, from https://www.kaggle.com/datasets/wordsforthewise/lending-club?resource=download. n.d.

[15] J. Yan. "Multivariate Modeling with Copulas and Engineering Applications". In: *Springer Handbook of Engineering Statistics*. Ed. by H. Pham. Springer, London, 2023. DOI: 10.1007/978-1-4471-7503-2_46.

[16] T. M. Yhip and B. M. D. Alagheband. *The Practice of Lending: A Guide to Credit Analysis and Credit Risk*. Germany: Springer International Publishing, 2020.