CopulaCenR: Copula based Regression Models for Bivariate Censored Data in R

by Tao Sun and Ying Ding

Abstract Bivariate time-to-event data frequently arise in research areas such as clinical trials and epidemiological studies, where the occurrence of two events are correlated. In many cases, the exact event times are unknown due to censoring. The copula model is a popular approach for modeling correlated bivariate censored data, in which the two marginal distributions and the between-margin dependence are modeled separately. This article presents the R package **CopulaCenR**, which is designed for modeling and testing bivariate data under right or (general) interval censoring in a regression setting. It provides a variety of Archimedean copula functions including a flexible two-parameter copula and different types of regression models (parametric and semiparametric) for marginal distributions. In particular, it implements a semiparametric transformation model for the margins with proportional hazards and proportional odds models being its special cases. The numerical optimization is based on a novel two-step algorithm. For the regression parameters, three likelihood-based tests (Wald, generalized score and likelihood ratio tests) are also provided. We use two real data examples to illustrate the key functions in **CopulaCenR**.

Introduction

Bivariate data arise frequently in many research areas such as health, epidemiology, and economics. For example, bivariate time-to-event endpoints are often used in clinical trials studying bilateral diseases (e.g., eye diseases) or complex diseases (e.g., cancer and psychiatric disorders). The two events are correlated as they come from the same subject. In many situations, the two event times cannot be precisely observed, leading to bivariate censored data. Specifically, bivariate right-censored data occur when the study ends prior to the occurrence of one or both events. An example of such data comes from a clinical study assessing the treatment effect on preventing blindness in Diabetic Retinopathy patients where each patient had one eye randomized to the treatment and the other eye received no treatment (Huster et al., 1989), and the time-to-blindness are bivariate and right-censored. We will illustrate the analysis of this study in Section 2.4. In another situation, bivariate interval-censored data occur when the status of both events are periodically examined at intermittent assessment times. In this case, the right censoring could also happen if the event still does not occur at the last assessment time. A special case of interval-censored data is the current status data if there is only one assessment time and the event is only known to occur or not by its assessment time. An example of bivariate interval-censored data will be demonstrated in Section 2.4, which came from a clinical trial studying the progression of a bilateral eye disease, Age-related Macular Degeneration (AMD), where the progression time to late-AMD are interval or right censored (AREDS Group, 1999). More examples can be found in books Hougaard (2000) and Sun (2007).

The development of our package is motivated by researches that are interested in (1) discovering covariates that are significantly associated with the bivariate censored outcomes, and (2) characterizing the joint and conditional risks of two events. For the bivariate events, the joint and conditional risks could be clinically more important than the marginal risk (of a single event). For example, the joint 5-year progression-free probability for both eyes helps identify patients with a high risk of progressing to late-AMD. For another example, for patients having one eye already progressed, the conditional 5-year progression-free probability for the non-progressed eye (given its fellow eye already progressed) provides important information for both clinicians and the patient since patients with both eyes progressed to the late stage of the disease may lose the ability to live independently.

There are three major approaches to fit regression models for bivariate censored data. The simplest way is to fit a marginal model and estimate the variance-covariance by a robust sandwich estimator (for example, Wei et al., 1989). This approach takes a working independence assumption, and thus cannot generate joint or conditional distributions. The second approach is based on frailty models (for example, Oakes, 1982), which are essentially mixed effects models and account for the dependence between two events by a latent frailty variable. However, the covariate effects in frailty models are usually interpreted on a conditional level (by conditioning on the frailty term), which is not straightforward. The third approach is to use copula models (for example, Clayton, 1978). Unlike the marginal or frailty approaches, the copula approach models the joint survival distribution by directly connecting the two marginal distributions through a copula function. One unique advantage of the copula is that it separately models the marginal distributions and the dependence parameter(s), allowing flexibility in marginal models and straightforward interpretation

for covariate effects. Moreover, the challenge from censoring can be naturally handled through the marginal distributions within the copula function. Besides, the joint and conditional distributions can be obtained based on the copula model.

Along with these three major approaches, multiple endeavors have been devoted to the development of software, mostly R (R Core Team, 2019) packages, to build regression models for bivariate censored data. For bivariate right-censored data, the **survival** (Therneau, 2018b) package can fit parametric or semiparametric Cox (Cox, 1972) marginal and frailty models. Also, packages such as **parfm** (Munda et al., 2012) and **frailtypack** (Rondeau et al., 2012) implement proportional hazards (PH) frailty models under the parametric and semiparametric settings. Other R packages such as **coxme** (Therneau, 2018a) and **phmm** (Donohue and Xu, 2019) also fit PH frailty models for right-censored data. For bivariate interval-censored data, the **survival** and **frailtypack** packages provide marginal and frailty models under the parametric or semiparametric (Cox PH) situation, respectively. The C++ program `IntCens` (codes located under `https://dlin.web.unc.edu/software/intcens/`) implements a class of semiparametric frailty models, including both PH and proportional odds (PO) models.

To the best of our knowledge, there exists no R package for fitting copula-based regression models for both bivariate right-censored and interval-censored data. The existing copula packages for bivariate data handle either the non-censoring (i.e., complete data) or the right-censoring situation. In the non-censoring situation, the package **copula** (Hofert et al., 2018) by Yan (2007) and Kojadinovic and Yan (2010) implements multivariate copula models without covariates for complete data and obtains the maximum likelihood estimator for the copula dependence parameter. It gives useful codes for implementing regression models in bivariate complete data in the appendix of Yan (2007). It also provides copula goodness-of-fit tests for model selection purpose. The package **VineCopula** (Schepsmeier et al., 2018) can also model bivariate or multivariate complete data without covariates through the vine copula models (Aas et al., 2009). Packages such as **CopulaRegression** (Nicole Kraemer, 2014) and **gcmr** (Masarotto and Varin, 2017) can provide copula-based regression models with parametric margins for bivariate or multivariate complete data and provide maximum likelihood estimators for model parameters. The package **gamCopula** (Nagler and Vatter, 2020) implements a generalized additive model that can take into account the effect of the predictors on the dependence structure of bivariate and vine copula models (Vatter and Chavez-Demoulin, 2015). For the right-censoring situation, the **Copula.surv** package (Emura, 2018) can estimate the Clayton copula dependence parameter in bivariate right-censored data without covariates and also perform a goodness-of-fit test for a fitted Clayton model (Emura et al., 2010). The **Sunclarco** package (Prenen et al., 2017b) provides Clayton or Gumbel copula-based regression models with parametric (Weibull and piecewise constant) or Cox semiparametric margins for multivariate right-censored data (Prenen et al., 2017a). The package **GJRM** (Marra and Radice, 2020) can fit both marginal and copula regression models in complete and right-censored data (Marra and Radice, 2017; Marra et al., 2017; Marra and Radice, 2019). By far, there is no copula-based R package for bivariate interval-censored data.

We develop the **CopulaCenR** package, which fits copula-based regression models for both bivariate right-censored and interval-censored data. The package is available from the Comprehensive R Archive Network (CRAN) at `https://CRAN.R-project.org/package=CopulaCenR`. The main advantage of **CopulaCenR** relies on the diverse choice of copula and marginal models for both bivariate right-censored and interval-censored data. Specifically, it provides a class of Archimedean copulas that correspond to a variety of dependence structures, as illustrated in Table 1. In particular, in addition to these frequently used one-parameter Archimedean copulas, a two-parameter copula function (Copula2) is also included. This Copula2 has more flexibility in modeling dependence structure, as we show in Section 2.2. Furthermore, **CopulaCenR** implements a list of parametric and semiparametric marginal regression models, as illustrated in Table 2. For parameter estimation, the package utilizes a novel two-step procedure that is computationally stable and efficient. For the inference of regression parameters, three likelihood-based tests such as Wald, generalized score and likelihood ratio tests are provided.

We will describe the major features of **CopulaCenR** in Section 2.2 and presents the model and estimation procedure in Section 2.3. We will demonstrate two real data examples in Section 2.4 using the version 1.1.2 of **CopulaCenR**. Finally, we will conclude and discuss in Section 2.5.

Package Features

The most popular copula family for bivariate censored data is the Archimedean copula family, which has an explicit form of

$$C_\eta(u,v) = \phi_\eta\{\phi_\eta^{-1}(u) + \phi_\eta^{-1}(v)\},$$

where u and v are two uniformly distributed margins; ϕ_η is the generator function, which is a continuous, strictly decreasing and convex function; ϕ_η^{-1} is the inverse of ϕ_η. One generator function uniquely

determines an Archimedean copula. The copula parameter η has a one-to-one correspondence with the popular dependence measure Kendall's τ. Another property of the copula is the tail dependence (i.e., τ_L and τ_U for lower and upper tail dependence), which measure the dependence between two margins in the lower and upper tails. More details about Archimedean copulas can be found in Nelsen (2006).

Table 1 lists six Archimedean family copula models that are implemented in **CopulaCenR**. Two most frequently used Archimedean copulas are Clayton (Clayton, 1978) and Gumbel (Gumbel, 1960) models, which account for the lower or upper tail dependence between two margins using a single parameter η. Other Archimedean copulas, such as Frank (Frank, 1979), Joe (Joe, 1993) and Ali-Mikhail-Haq (AMH) (Ali et al., 1978), are also one-parameter copulas. In addition to these five copulas, we also include a flexible two-parameter Archimedean copula model (Joe and Hu, 1996; Joe, 1997), namely, Copula2 (also called the "BB1" family), which is formulated as

$$C_{\alpha,\kappa}(u,v) = [1 + \{(u^{-1/\kappa} - 1)^{1/\alpha} + (v^{-1/\kappa} - 1)^{1/\alpha}\}^{\alpha}]^{-\kappa},\ \alpha \in (0,1],\ \kappa \in (0,\infty). \tag{1}$$

The two dependence parameters (α and κ) are explicitly connected to Kendall's τ with $\tau = 1 - 2\alpha\kappa/(2\kappa+1)$, and they account for the correlation between u and v at upper and lower tails. In particular, when $\alpha = 1$, Copula2 becomes the Clayton copula, and when $\kappa \to \infty$, it becomes the Gumbel copula. Thus, the two-parameter copula model provides more flexibility in modeling the between-margin dependence than the one-parameter copulas such as Clayton or Gumbel (Joe, 2014). Figure 1 illustrates the scatter plots of bivariate event times generated from the six copula models in Table 1.

Family	Parameter Space	Generator $\phi_\eta(t), t \in [0,\infty)$	Generator Inverse $\phi_\eta^{-1}(s), s \in (0,1]$	τ_L	τ_U	Kendall's τ
Clayton	$\eta > 0$	$(1+t)^{-1/\eta}$	$s^{-\eta} - 1$	$2^{-1/\eta}$	0	$\eta/(2+\eta)$
Gumbel	$\eta \geq 1$	$\exp(-t^{1/\eta})$	$(-\log s)^{\eta}$	0	$2 - 2^{1/\eta}$	$1 - 1/\eta$
Frank	$\eta \geq 0$	$-\eta^{-1}\log\{1 + e^{-t}(e^{-\eta} - 1)\}$	$-\log\{(e^{-\eta s} - 1)/(e^{-\eta} - 1)\}$	0	0	$1 + 4\{D_1(\eta) - 1\}/\eta$
AMH	$\eta \in [0,1)$	$(1-\eta)/(e^t - \eta)$	$\log[\{1 + \eta(s-1)\}/s]$	0	0	$1 - 2\{(1-\eta)^2\log(1-\eta) + \eta\}/(3\eta^2)$
Joe	$\eta \geq 1$	$1 - (1 - e^{-t})^{1/\eta}$	$-\log\{1 - (1-s)^{\eta}\}$	0	$2 - 2^{1/\eta}$	$1 - 4\sum_{k=1}^{\infty} 1/\{k(\eta k + 2)[\eta(k-1) + 2]\}$
Copula2	$\alpha \in (0,1], \kappa > 0$	$\{1/(1 + t^{\alpha})\}^{\kappa}$	$(s^{-1/\kappa} - 1)^{1/\alpha}$	$2^{-\alpha\kappa}$	$2 - 2^{\alpha}$	$1 - 2\alpha\kappa/(2\kappa + 1)$

τ_L and τ_U are the lower and upper tail dependence measures.
$D_1(\cdot)$ is the Debye function written as $D_1(\eta) = \frac{1}{\eta}\int_0^{\eta} \frac{t}{e^t - 1}\,dt$.

Table 1: Summary of implemented Archimedean copula families.

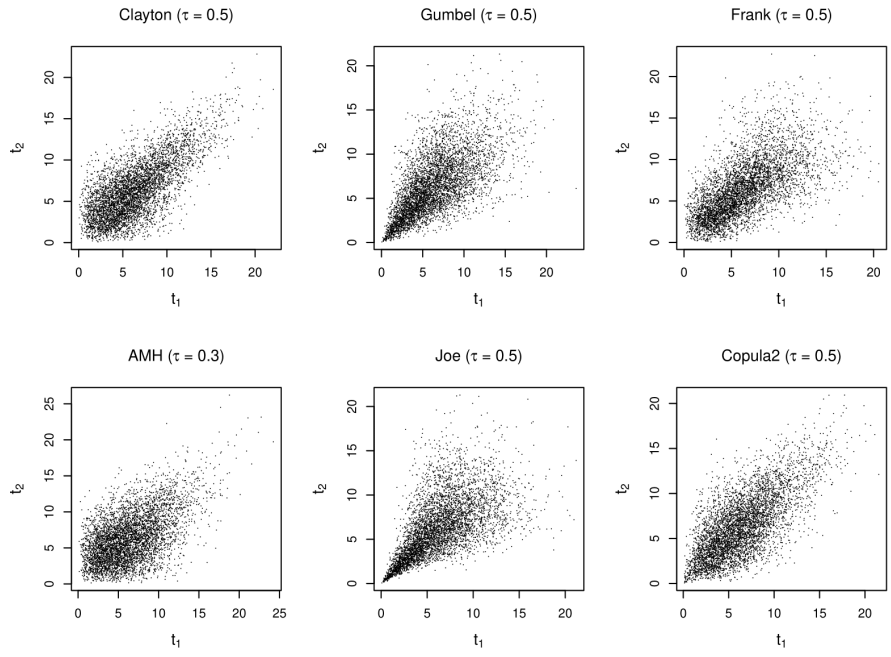

Figure 1: Scatter plots of bivariate event times generated from various copula models.

To fit a copula-based regression model, one also needs to choose a regression model for the margins. Table 2 lists the available marginal models in **CopulaCenR**. For bivariate right-censored data, users can fit either a parametric marginal model via the function `rc_par_copula` or a semiparametric Cox PH model via the function `rc_spCox_copula` (Sun et al., 2019). Specifically, the parametric models incorporate both the PH (e.g., Weibull, Gompertz) and the PO (e.g., Loglogistic) models. For

bivariate interval-censored data, one can choose to fit a parametric marginal model via the function `ic_par_copula`. Moreover, the package can also fit a semiparametric transformation model via the function `ic_spTran_copula`. It contains a variety of marginal models including the PH and PO models, as we explain in Section 2.3.3. A novel two-step sieve estimation procedure is implemented (Sun and Ding, 2019).

Type	Models	Survival Distributions $S(t)$	Corresponding R Functions
Parametric	Weibull	$\exp\{-(t/\lambda)^k e^{Z^\top \beta}\}$	`rc_par_copula`, `ic_par_copula`
	Gompertz	$\exp\{-\frac{b}{a}(e^{at}-1)e^{Z^\top \beta}\}$	
	Loglogistic	$\{1+(t/\lambda)^k e^{Z^\top \beta}\}^{-1}$	
Semiparametric	Cox	$\exp\{-\Lambda(t)e^{Z^\top \beta}\}$	`rc_spCox_copula`
	Transformation	$\exp[-G\{\Lambda(t)e^{Z^\top \beta}\}]$	`ic_spTran_copula`

Table 2: Summary of implemented marginal models.

For the inference of the covariate effects, three types of likelihood-based tests are implemented in **CopulaCenR**: the Wald test (built within `rc_par_copula`, `rc_spCox_copula`, `ic_par_copula`, and `ic_spTran_copula`), the generalized score test (`score_copula`) and the likelihood-ratio test (`lrt_copula`).

After a copula model being fitted, fitted values (i.e., linear predictors, survival probabilities) can be extracted by the general S3 function `fitted`. For new observations, the linear predictors and survival probabilities can be obtained using the function `predict`. Moreover, the user can plot three types of distributions (joint, conditional and marginal) using the general functions `plot` and `lines`. In particular, an interactive 3D contour will be plotted to visualize the joint distribution.

Besides, the package provides a bivariate event time generating function `data_sim_copula`, which can generate random bivariate event times based on a specified copula function, a marginal distribution, and covariate values.

In summary, the key functions of **CopulaCenR** are

- `rc_par_copula`: for fitting parametric regression models to bivariate right-censored data;
- `rc_spCox_copula`: for fitting a semiparametric Cox regression model to bivariate right-censored data;
- `ic_par_copula`: for fitting parametric regression models to bivariate interval-censored data;
- `ic_spTran_copula`: for fitting a semiparametric transformation model to bivariate interval-censored data;
- `score_copula`: for performing the generalized score test on covariate effects;
- `lrt_copula`: for performing the likelihood ratio test (LRT) on covariate effects between two nested models;
- `tau_copula`: for calculating Kendall's τ from copula parameter estimates;
- `plot,lines`: S3 methods for plotting joint, conditional and marginal distributions based on a fitted copula model;
- `fitted,predict`: S3 methods for extracting fitted values and predicting new observations;
- `summary,print,coef,logLik,AIC,BIC`: other S3 functions for a fitted object;
- `data_sim_copula`: for generating bivariate event times through a specified copula model and marginal distributions.

We use two real data examples to illustrate the implementation of these functions in Section 2.4.

Methods

Copula model for bivariate censored data

Let (T_1, T_2) be the true bivariate event times, with marginal survival functions $S_j(t_j) = Pr(T_j > t_j)$, $j = 1, 2$, and joint survival function $S(t_1, t_2) = Pr(T_1 > t_1, T_2 > t_2)$. Assume there are n independent subjects in a study. When (T_1, T_2) are subject to right-censoring, for subject $i = 1 \cdots n$, we observe $D_i = \{(Y_{ij}, \Delta_{ij}, Z_{ij}) : Y_{ij} = \min(T_{ij}, C_{ij}), \Delta_{ij} = I(T_{ij} \leq C_{ij}), j = 1, 2\}$, where C_{ij} is the censoring time of T_{ij}, Δ_{ij} is the censoring indicator and Z_{ij} is the covariate vector. When (T_1, T_2) are under interval-censoring, we observe $D_i = \{(L_{ij}, R_{ij}, Z_{ij}), j = 1, 2\}$ for subject i, where $(L_{ij}, R_{ij}]$ is the time interval that T_{ij} lies in and Z_{ij} is the covariate vector.

By the Sklar's theorem (Sklar, 1959), so long as the marginal survival functions S_j are continuous, there exists a unique function C_η that connects two marginal survival functions into the joint survival function: $S(t_1,t_2) = C_\eta\{S_1(t_1), S_2(t_2)\},\ t_1,t_2 \geq 0$. Here, the function C_η is called a copula and its parameter η measures the dependence between the two margins. A signature feature of the copula is that it allows the dependence to be modeled separately from the marginal distributions.

Joint likelihood functions for bivariate censored data

In this section, we present the joint likelihood functions for bivariate right-censored data and bivariate interval-censored data, respectively.

Define the density function for copula $C_\eta(u,v)$ as $c_\eta(u,v) = \partial^2 C_\eta(u,v)/\partial u \partial v$. Let $f(t_1,t_2) = \partial^2 S(t_1,t_2)/\partial t_1 \partial t_2 = c_\eta\{S_1(t_1), S_2(t_2)\} f_1(t_1) f_2(t_2)$ denote the corresponding density function of $S(t_1,t_2)$. Denote by $\theta = (\beta^\top = (\beta_1^\top, \beta_2^\top), \eta, S_{01}, S_{02})^\top$ all the unknown parameters in $S(t_1,t_2)$, where β_j is the regression coefficient vector and S_{0j} is the baseline survival function for the jth margin. Then, the joint likelihood for the observed data $D = \{D_i\}_{i=1}^n$ can be written as

$$
\begin{aligned}
L_n(\theta|D) =& \prod_{i=1}^n f(y_{i1}, y_{i2}|Z_{i1}, Z_{i2})^{\delta_{i1}\delta_{i2}} \times \left[-\frac{\partial S(y_{i1}, y_{i2}|Z_{i1}, Z_{i2})}{\partial y_{i1}}\right]^{\delta_{i1}(1-\delta_{i2})} \\
&\times \left[-\frac{\partial S(y_{i1}, y_{i2}|Z_{i1}, Z_{i2})}{\partial y_{i2}}\right]^{(1-\delta_{i1})\delta_{i2}} \times S(y_{i1}, y_{i2}|Z_{i1}, Z_{i2})^{(1-\delta_{i1})(1-\delta_{i2})} \\
=& \prod_{i=1}^n \left[c_\eta\{S_1(y_{i1}|Z_{i1}), S_2(y_{i2}|Z_{i2})\} f_1(y_{i1}|Z_{i1}) f_2(y_{i2}|Z_{i2})\right]^{\delta_{i1}\delta_{i2}} \\
&\times \left[-\frac{\partial C_\eta\{S_1(y_{i1}|Z_{i1}), S_2(y_{i2}|Z_{i2})\}}{\partial y_{i1}}\right]^{\delta_{i1}(1-\delta_{i2})} \\
&\times \left[-\frac{\partial C_\eta\{S_1(y_{i1}|Z_{i1}), S_2(y_{i2}|Z_{i2})\}}{\partial y_{i2}}\right]^{(1-\delta_{i1})\delta_{i2}} \\
&\times C_\eta\{S_1(y_{i1}|Z_{i1}), S_2(y_{i2}|Z_{i2})\}^{(1-\delta_{i1})(1-\delta_{i2})},
\end{aligned}
\tag{2}
$$

where $(\delta_{i1}, \delta_{i2}) \in \{(0,0), (0,1), (1,0), (1,1)\}$.

Similarly, using the notation introduced in Section 2.3.1, the joint likelihood function for bivariate interval-censored data from n subjects can be written as

$$
\begin{aligned}
L_n(\theta|D) &= \prod_{i=1}^n Pr(L_{i1} < T_{i1} \leq R_{i1}, L_{i2} < T_{i2} \leq R_{i2}|Z_{i1}, Z_{i2}) \\
&= \prod_{i=1}^n \Big[Pr(T_{i1} > L_{i1}, T_{i2} > L_{i2}|Z_{i1}, Z_{i2}) - Pr(T_{i1} > L_{i1}, T_{i2} > R_{i2}|Z_{i1}, Z_{i2}) \\
&\qquad - Pr(T_{i1} > R_{i1}, T_{i2} > L_{i2}|Z_{i1}, Z_{i2}) + Pr(T_{i1} > R_{i1}, T_{i2} > R_{i2}|Z_{i1}, Z_{i2}) \Big] \\
&= \prod_{i=1}^n \Big[C_\eta\{S_1(L_{i1}|Z_{i1}), S_2(L_{i2}|Z_{i2})\} - C_\eta\{S_1(L_{i1}|Z_{i1}), S_2(R_{i2}|Z_{i2})\} \\
&\qquad - C_\eta\{S_1(R_{i1}|Z_{i1}), S_2(L_{i2}|Z_{i2})\} + C_\eta\{S_1(R_{i1}|Z_{i1}), S_2(R_{i2}|Z_{i2})\} \Big].
\end{aligned}
\tag{3}
$$

The right interval R_{ij} can take values in $(0,\infty]$. For a given subject i, if $R_{ij} = \infty$ (i.e., T_{ij} is right-censored), then any term involving R_{ij} becomes 0 and the joint survival function for subject i reduces to only one (if both R_{i1} and R_{i2} are ∞) or two (if one R_{ij} is ∞) terms. The special case of bivariate current status data (i.e., only one assessment time for each subject) can also fit into this framework, where for each T_{ij}, either $L_{ij} = 0$ (T_{ij} is smaller than the assessment time, which is R_{ij} in this case) or $R_{ij} = \infty$ (T_{ij} is greater than the assessment time, which is L_{ij} in this case). Therefore, the likelihood function (3) can handle the bivariate data under general interval-censoring.

Marginal models

We implement several popular parametric marginal models in **CopulaCenR**, as shown in Table 2. For example, the marginal Weibull survival distribution can be written as

$$
S_j(t_j|Z_j) = \exp\{-(t_j/\lambda_j)^{k_j} e^{Z_j^\top \beta_j}\},\ j = 1,2,
$$

where λ_j and k_j are the scale and shape parameters of the baseline Weibull distribution, and β_j are the covariate effects. The model follows the PH assumption. In this case, the parameter set θ becomes $(\beta^\top, \eta, \lambda_1, k_1, \lambda_2, k_2)^\top$. Other parametric distributions including Gompertz and Loglogistic are also implemented in the package.

More generally, we implement the semiparametric Cox PH marginal model for bivariate right-censored data. The model does not specify the marginal distribution for the baseline hazards function. Instead, the baseline hazards are treated as piecewise constants between all uncensored event times as suggested by Breslow (1972). The model is expressed as

$$S_j(t_j|Z_j) = \exp\{-\Lambda_j(t_j)e^{Z_j^\top \beta_j}\},\ j = 1, 2,$$

in which the Breslow baseline cumulative hazard function $\Lambda_j(t)$ is given by

$$\Lambda_j(t) = \sum_{i=1}^{n} \frac{I(Y_{ij} \leq t)\delta_{ij}}{\sum_{k \in R_{ij}} \exp Z_k^\top \beta_j},$$

where $R_{ij} = \{k : Y_k \geq Y_{ij}\}$ denotes the at-risk set at time Y_{ij}.

We also consider a class of semiparametric linear transformation models for the marginal distribution of the interval-censored data. The model is expressed as:

$$S_j(t_j|Z_j) = \exp[-G_j\{\Lambda_j(t_j)e^{Z_j^\top \beta_j}\}],\ j = 1, 2. \tag{4}$$

$\Lambda_j(\cdot)$ is an unknown and non-decreasing function of t, which is not necessarily the baseline cumulative hazards function. In **CopulaCenR**, we approximate Λ_j in a sieve space constructed by Bernstein polynomials. A Bernstein basis polynomial with degree m is expressed as:

$$B_k(t, m, l, u) = \binom{m}{k} \left(\frac{t-l}{u-l}\right)^k \left(1 - \frac{t-l}{u-l}\right)^{m-k},\ k = 0, ..., m, \tag{5}$$

where l and u are the lower and upper bounds of all observed times. One big advantage of Bernstein polynomials is that they do not require the specification of interior knots, as seen from (5), making them easy to work with. More details can be found in Sun and Ding (2019).

In model (4), $G_j(\cdot)$ is a pre-specified strictly increasing function, such as the Box-Cox and the logarithmic transformation functions. The package uses a $G(\cdot)$ function as specified in Zhou et al. (2017):

$$G_j(x) = \begin{cases} \frac{(1+x)^r - 1}{r},\ 0 < r \leq 2, \\ \frac{\log\{1+(r-2)x\}}{r-2},\ r > 2. \end{cases} \tag{6}$$

Note that the model (4) contains a class of survival models. For example, when $G(x) = x$ at $r = 1$, the marginal function $S_j(t|Z)$ becomes $\exp\{-\Lambda_j(t)e^{Z^\top \beta_j}\}$, which is essentially a PH model. Likewise, when $G_j(x) = \log(1 + x)$ at $r = 3$, $S_j(t|Z)$ becomes $\{1 + \Lambda_j(t)e^{Z^\top \beta_j}\}^{-1}$, which is a PO model. In practice, the value of r can be either selected according to model AIC or treated unknown and estimated together with other model parameters.

Two-step estimation procedure

In this section, we illustrate the estimation procedure for the unknown parameter θ. For simplicity, we use the general notation $\theta = (\beta_1^\top, \beta_2^\top, \eta, S_{01}, S_{02})^\top$ throughout this section. In principle, we can maximize the joint log-likelihood function based on formula (2) or (3) directly, written as $l_n(\theta|D) = \log L_n(\theta|D) = \sum_{i=1}^n \log L(\theta|D_i)$. Due to the complex structure of the log-likelihood function, we implement a novel two-step estimation procedure, which is proven to be computationally more stable and efficient than the one-step procedure, as shown in Sun et al. (2019) and Sun and Ding (2019). Essentially, the two-step procedure implements an extra step to obtain appropriate initial values for all the unknown parameters. The estimation procedure is described below:

1. Obtain initial estimates of θ_n:
   (a) $(\hat{\beta}_{jn}^{(1)}, \hat{S}_{0j}^{(1)}) = \underset{(\beta_j, S_{0j})}{\arg\max}\, l_{jn}(\beta_j, S_{0j})$, where l_{jn} denotes the log-likelihood for the marginal model, $j = 1, 2$;

(b) $\hat{\eta}_n^{(1)} = \arg\max_{(\eta)} l_n\{\hat{\beta}_n^{(1)} = (\hat{\beta}_{1n}^{(1)}, \hat{\beta}_{2n}^{(1)}), \eta, \hat{S}_{01}^{(1)}, \hat{S}_{02}^{(1)}\}$, where $\hat{\beta}_{jn}^{(1)}$ and $\hat{S}_{0j}^{(1)}$ are the initial estimates from (a), and l_n is the joint log-likelihood.

2. Simultaneously maximize the joint log-likelihood to get final estimates: $\hat{\theta}_n = (\hat{\beta}_n, \hat{\eta}_n, \hat{S}_{01}, \hat{S}_{02}) = \arg\max_{(\beta,\eta,S_{01},S_{02})} l_n(\beta, \eta, S_{01}, S_{02})$ with initial values $(\hat{\beta}_n^{(1)}, \hat{\eta}_n^{(1)}, \hat{S}_{01}^{(1)}, \hat{S}_{02}^{(1)})$ obtained from step 1(a) and 1(b).

[Remark 1.] In the case of semiparametric Cox PH margins (with the Breslow baseline cumulative hazard estimator), although the maximum likelihood estimators from step 2 are consistent and asymptotically normal, the Hessian matrix cannot be directly used for estimating the variance-covariance matrix of $(\hat{\beta}, \hat{\eta})$ (Sun et al., 2019). Therefore, the bootstrap procedure is implemented in the package for producing a valid variance-covariance estimator.

[Remark 2.] In the case of semiparametric transformation model margins (with the use of Bernstein polynomials), the two-step estimation procedure becomes a two-step "sieve" estimation procedure. Sun and Ding (2019) rigorously proved the asymptotic properties of the sieve maximum likelihood estimators.

The main model-fitting functions (`rc_par_copula`, `rc_spCox_copula`, `ic_par_copula` and `ic_spTran_copula`) provide a built-in optimization option, which is a wrapper to the optimization routines `optim` and `nlm` in R.

Likelihood-based tests for covariate effects

We now separate β into two parts: β_g and β_{ng}, where β_g is the parameter set of interest for hypothesis testing and β_{ng} denotes the rest of the regression coefficients. In certain cases, β_g can be the entire regression parameter β. The package implements three likelihood-based tests including the Wald test, the generalized score test (Cox and Hinkley, 1979) and the likelihood ratio test, which are asymptotically equivalent and follow the chi-squared distribution with $df = \dim(\beta_g)$. In particular, the generalized score test is usually faster than the other two tests for large-scale testings such as the genome-wide association study (GWAS) (Sun et al., 2019; Sun and Ding, 2019). Due to the complex structure of the joint log-likelihood, instead of analytically deriving the first and second order derivatives, we use the Richardson's extrapolation (Lindfield et al., 1989) to approximate the score function and observed Fisher information numerically.

Examples

Bivariate event time generation

The package **CopulaCenR** provides a user-friendly function `data_sim_copula` for generating random bivariate event times based on a specified copula model, marginal distributions and covariate values. The arguments `n`, `copula`, and `eta` assign the sample size, the copula type, and the dependence parameter value. For marginal distributions, the argument `dist` can be one of the three parametric distributions in Table 2 (i.e., Weibull, Loglogistic and Gompertz), and their distribution parameters are given through `baseline`. For Weibull and Loglogistic, the baseline parameters are λ (scale) and k (shape); whereas a (shape) and b (rate) for the Gompertz distribution. In this current version, we assume that the two margins share the same set of covariates and effects, which are assigned by `var_list` and `COV_beta`, respectively. Lastly, `x1` and `x2` input a data frame of covariate values for the two margins, respectively. Figure 2 illustrates a scatter plot of 500 simulated bivariate event times from a Clayton model with Weibull margins, as demonstrated in the code below.

```
library(CopulaCenR)
set.seed(1)
dat <- data_sim_copula(n = 500, copula = "Clayton", eta = 3, dist = "Weibull",
                baseline = c(0.1,2), var_list = c("var1", "var2"), COV_beta = c(0.1, 0.1),
                x1 = cbind(rnorm(500, 6, 2), rbinom(500, 1, 0.5)),
                x2 =  cbind(rnorm(500, 6, 2), rbinom(500, 1, 0.5)))

head(dat)

id ind      var1 var2     time
 1   1  6.130533    1 8.062168
 1   2  6.154606    1 7.472649
```

```
 2   1  8.070653    1 6.317247
 2   2  5.406263    1 5.904064
 3   1 10.520432    1 4.195788
 3   2  3.633516    1 4.771523
```

```
plot(x = dat$time[dat$ind == 1], y = dat$time[dat$ind == 2],
     xlab = expression(t[1]), ylab = expression(t[2]), cex.axis = 1, cex.lab = 1.3)
```

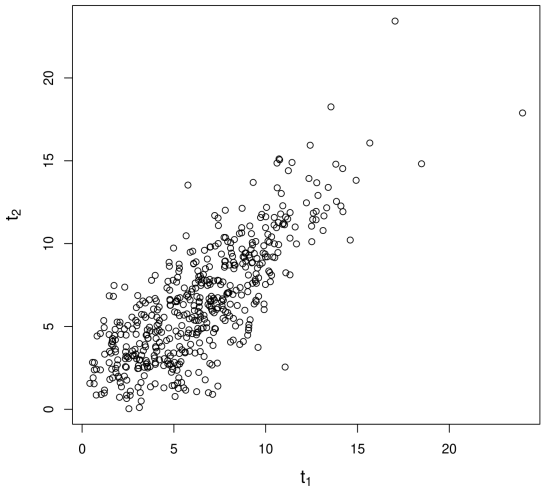

Figure 2: Simulated bivariate event times from the Clayton copula with Weibull margins.

Fitting copula models for bivariate right-censored data

The bivariate right-censored input dataset shall be a data frame including the covariates and four additional key input columns:

- id: the subject/cluster id,
- ind: the margin indicator (1, 2),
- obs_time: the exact observed time,
- status: censoring indicator (1 for event, 0 for right-censoring).

We use the DRS (Diabetic Retinopathy Study) data as an example. The DRS data contain bivariate right-censored time to blindness from 197 diabetic retinopathy patients. These patients were from a 50% random sample of the patients with "high-risk" diabetic retinopathy as defined by the DRS (Huster et al., 1989). Each patient had one eye randomized to one of the two laser treatments and the other eye received no treatment. For each eye, the event of interest was the time from initiation of treatment to the time to blindness in months. Censoring was caused by death, dropout, or end of the study. The data can be loaded by

```
data("DRS", package = "CopulaCenR")
head(DRS)
```

```
id  ind obs_time status treat age type
 5    1    46.23      0     0  28    2
 5    2    46.23      0     2  28    2
14    1    42.50      0     2  12    1
14    2    31.30      1     0  12    1
16    1    42.27      0     1   9    1
16    2    42.27      0     0   9    1
```

There are three covariates: `treat` is treatment with 0 for no treatment, 1 for xenon laser treatment and 2 for argon laser treatment; `age` is the age at diagnosis of diabetes; `type` is the type of diabetes

with 1 for juvenile (age ≤ 20 at diagnosis) and 2 for adult. The primary question of the DRS study was to assess the treatment effectiveness while accommodating the dependence between two eyes.

We now demonstrate how to fit a Clayton copula model with Weibull margins to the DRS data using the function `rc_par_copula`. We are interested in the treatment effect, as indicated in argument `var_list`. The arguments `copula` and `m.dis` specify the fitted copula model and marginal baseline distributions. The default optimization method is `BFGS` (Nash, 1990). Other optimization methods and control parameters can also be applied (see `?optim`).

```
library(CopulaCenR)
clayton_wb <- rc_par_copula(data = DRS, var_list = "treat", copula = "Clayton",
                            m.dist = "Weibull", method = "BFGS")

summary(clayton_wb)

Copula:   Clayton
Margin:   Weibull

         estimate         SE     stat    pvalue
lambda 90.6440318 13.1887218  47.2360 6.293e-12 ***
k       0.8062766  0.0586207 189.1758 < 2.2e-16 ***
treat1 -0.5714498  0.1997080   8.1878  0.004217 **
treat2  0.0052997  0.1739106   0.0009  0.975689
eta     0.6205855  0.2610638   5.6508  0.017447 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(The Wald tests are testing whether each coefficient is 0)

Final llk:   -839.7212
Convergence is completed successfully
```

The estimation and Wald test results suggest the xenon treatment significantly reduced the risk of blindness compared to controls ($p = 0.004217$ for `treat1`). We also compared our estimates with previous findings in Huster et al. (1989). Due to the differences in the model parameterization, we first transformed our estimates into comparable forms. Specifically, the xenon treatment effect in Huster et al. (1989) can be expressed as $-k\log(\lambda) + treat1 = -4.20$ and similarly the argon treatment effect is $-k\log(\lambda) + treat2 = -3.63$, which are consistent with the reported estimates $(-4.20, -3.42)$ in the Table 2 (page 151) of Huster et al. (1989). The AIC and BIC of this model can be obtained from the S3 methods `AIC` and `BIC`.

```
AIC(clayton_wb)

1689.442

BIC(clayton_wb)

1705.858
```

After the model is fitted, Kendall's τ can be estimated through the function `tau_copula`.

```
tau_copula(eta = as.numeric(coef(clayton_wb)["eta"]), copula = "Clayton")

0.2368118
```

The fitted values (i.e., linear predictors and survival probabilities) can be extracted through the function `fitted`. As the model is a PH model, the linear predictors (`type` is "lp") are the estimated log proportional hazards.

```
fit1 <- fitted(clayton_wb, type = "lp")
fit1[1:3, ]

id          lp1         lp2
 5  0.000000000 0.005299655
14  0.005299655 0.000000000
16 -0.571449835 0.000000000
```

When `type` is "survival", the fitted outputs are marginal (`S1`, `S2`) and joint (`S12`) survival probabilities at the observed times (`t1`, `t2`).

```
fit2 <- fitted(clayton_wb, type = "survival")
fit2[1:3, ]

id    t1    t2        S1        S2       S12
 5 46.23 46.23 0.5592967 0.5575724 0.3643588
14 42.50 31.30 0.5793467 0.6542323 0.4234880
16 42.27 42.27 0.7369175 0.5823995 0.4655204
```

Similarly, the `predict` function provides predictions for new observations with covariates. Its outputs can be either linear predictors or survival probabilities (at specified times). The following `newdata1` example contains two subjects under different treatments.

```
newdata1 <- data.frame(id = rep(1:2, each=2), ind = rep(c(1,2),2),
                       time = rep(40,4), treat = factor(c(0,1,0,2)))
newdata1

id ind time treat
 1   1   40     0
 1   2   40     1
 2   1   40     0
 2   2   40     2

predict(clayton_wb, newdata = newdata1, type = "lp")

id lp1          lp2
 1   0 -0.571449835
 2   0  0.005299655

predict(clayton_wb, newdata = newdata1, type = "survival")

id t1 t2        S1        S2       S12
 1 40 40 0.5962669 0.7467754 0.4799705
 2 40 40 0.5962669 0.5946309 0.4024998
```

Fitting copula models for bivariate interval-censored data

The bivariate interval-censored input dataset shall be a data frame including the covariates and five key input columns:

- id: the subject/cluster id,
- ind: the margin indicator (1 or 2),
- Left: the left bound of the observed interval,
- Right: the right bound of the observed interval (can take "Inf"),
- status: the censoring indicator (1 for left- or interval-censoring, 0 for right-censoring).

We use the AREDS (Age-Related Eye Disease Study) data as an example. The event of interest is the AMD (Age-related Macular Degeneration) disease, which is a leading cause of blindness in the developed world (Swaroop et al., 2009). It is known as a polygenic, progressive and neurodegenerative disorder. The AREDS study is a multi-center randomized clinical trial studying the development and progression of AMD, sponsored by the National Eye Institute (AREDS Group, 1999). Due to intermittent assessment times (every 6 months up to the first 6 years and every 1 year since after), the exact time when each eye progressed to late-AMD was only known to lie in a certain interval. As a result, the outcome data are bivariate interval-censored. The package includes a subset data of 629 Caucasian participants from AREDS who had at least one eye in moderate AMD stage at baseline. The data can be loaded by

```
data("AREDS", package = "CopulaCenR")
head(AREDS)

id ind Left Right status SevScaleBL ENROLLAGE rs2284665
 1   1  0.0   2.0      1          6      67.0         1
 1   2  0.0   2.0      1          8      67.0         1
 2   1  0.0   2.0      1          7      68.0         0
 2   2  5.9   9.3      1          4      68.0         0
```

```
 3   1  8.0   9.1      1          7      64.9         0
 3   2 10.0   Inf      0          7      64.9         0
```

Out of these 629 subjects, 273 subjects developed late-AMD in both eyes during the study and the times to late-AMD were interval-censored; 138 subjects developed late-AMD in one eye (interval-censored) and did not develop late-AMD before the end of the study (right-censored); the rest 218 subjects were right-censored for late-AMD in both eyes.

There are three continuous covariates: `SevScaleBL` for baseline AMD severity score (a value between 1 and 8 with a higher value indicating more severe AMD), `ENROLLAGE` for baseline age and `rs2284665` for a genetic variant ($0, 1, 2$ for GG, GT, TT) that might be associated with AMD progression. The two clinical covariates `SevScaleBL` and `ENROLLAGE` are well-known risk factors of AMD. Thus, our primary interest is to find out whether the genetic variant `rs2284665` is significantly associated with AMD progression.

We fit a two-parameter copula semiparametric transformation model for the AREDS data through the function `ic_spTran_copula`. The arguments `l` and `u` are the range of event times, which need to be pre-specified by the user. In practice, `l` and `u` can be set as the minimum and maximum of observed times. The argument `m` corresponds to the degree of Bernstein polynomials (as shown in formula 5), with the default value $m = 3$. The argument `r` specifies the form of marginal transformation model (as shown in formula 6). In practice, the values of `m` and `r` can be chosen based on the smallest AIC for a list of fitted models with different values.

We now demonstrate how to fit a two-parameter copula semiparametric model to the AREDS data. We chose the range of event times as $l = 0$ and $u = 15$, use the default Bernstein polynomial degree as $m = 3$ and assume PO for the margins (i.e., $r = 3$).

```
library(CopulaCenR)
copula2_sp <- ic_spTran_copula(data = AREDS, copula = "Copula2",
                               var_list = c("ENROLLAGE","rs2284665","SevScaleBL"),
                               l = 0, u = 15, m = 3, r = 3)
summary(copula2_sp)

Copula:   Copula2
Margin:  semiparametric

           estimate       SE    stat    pvalue
ENROLLAGE  0.042610 0.012271  12.057 0.0005159 ***
rs2284665  0.397712 0.091180  19.026 1.290e-05 ***
SevScaleBL 0.722681 0.053258 184.132 < 2.2e-16 ***
alpha      0.930508 0.058714 251.167 < 2.2e-16 ***
kappa      0.974037 0.226081  18.562 1.645e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(The Wald tests are testing whether each coefficient is 0)

Final llk:   -2104.178
Convergence is completed successfully
```

From the output, the estimated odds ratio for the genetic variant *rs*2284665 is $\exp(0.397712) = 1.49$ with a p-value 1.29×10^{-5}, implying it has a "harmful" effect for AMD patients by having more copies of its minor allele T. The AIC and BIC values are 4226.356 and 4266.353, respectively.

```
AIC(copula2_sp)

4226.356

BIC(copula2_sp)

4266.353
```

Also, the estimated Kendall's τ is 0.38, suggesting moderate dependence in AMD progression between two eyes.

```
tau_copula(eta = as.numeric(coef(copula2_sp)[c("alpha","kappa")]),
           copula = "Copula2")

0.3851248
```

Furthermore, we can test the effect of *rs*2284665 by the generalized score test. We first fit a null model without *rs*2284665 and then test its effect using the function `score_copula`.

```
copula2_sp_null <- ic_spTran_copula(data = AREDS, copula = "Copula2",
                                    var_list = c("ENROLLAGE","SevScaleBL"),
                                    l = 0, u = 15, m = 3, r = 3)
score_copula(object = copula2_sp_null, var_score = "rs2284665")

        stat       pvalue
1.943163e+01 1.042661e-05
```

The LRT can also be performed by applying two nested models to the function `lrt_copula`.

```
lrt_copula(model1 = copula2_sp, model2 = copula2_sp_null)

       stat      pvalue
9.543119588 0.002007003
```

The following codes plot the 3D joint survival probabilities for the three subjects in `newdata2`, which have the same `SevScaleBL = 3` in both eyes and `ENROLLAGE = 60`, but vary in the genotype of `rs2284665`. In the `plot` function, the argument `class` specifies the plot type, which can be one of "joint", "conditional" and "marginal". When `class = "joint"`, it generates a 3D interactive contour that can be manually rotated for the desired visualization. Figure 3 is a snapshot of 3D contours for the three subjects in `newdata2`.

```
newdata2 <- data.frame(id = rep(1:3, each=2), ind = rep(c(1,2),3),
                       SevScaleBL = rep(3,6), ENROLLAGE = rep(60,6),
                       rs2284665 = c(0,0,1,1,2,2))
newdata2

id ind SevScaleBL ENROLLAGE rs2284665
 1   1          3        60         0
 1   2          3        60         0
 2   1          3        60         1
 2   2          3        60         1
 3   1          3        60         2
 3   2          3        60         2

plot(x = copula2_sp, class = "joint", newdata = newdata2)
```

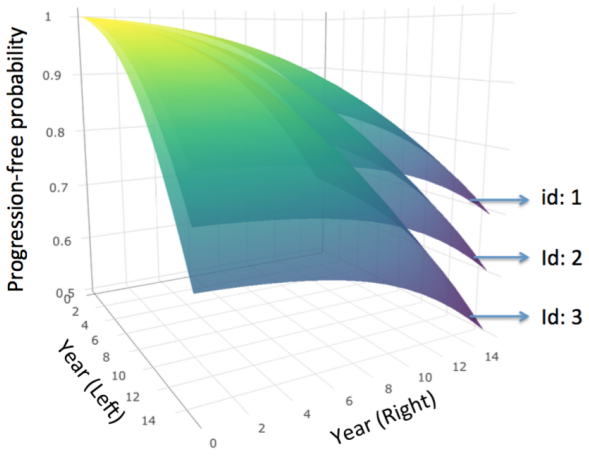

Figure 3: Estimated joint progression-free probability contours for subjects with different genotypes of *rs*2284665 (age 60 and severity score 3 in both eyes).

Similarly, the conditional survival probabilities (Figure 4) can be obtained for the left eyes from the same three subjects, given their right eyes (i.e., `cond_margin` = 2) had progressed (to late-AMD) at year 5 (i.e., `cond_time` = 5).

```
plot(x = copula2_sp, class = "conditional", newdata = newdata2,
     cond_margin = 2, cond_time = 5, ylim = c(0.25,1),
     ylab = "Conditional Survival Probability")
```

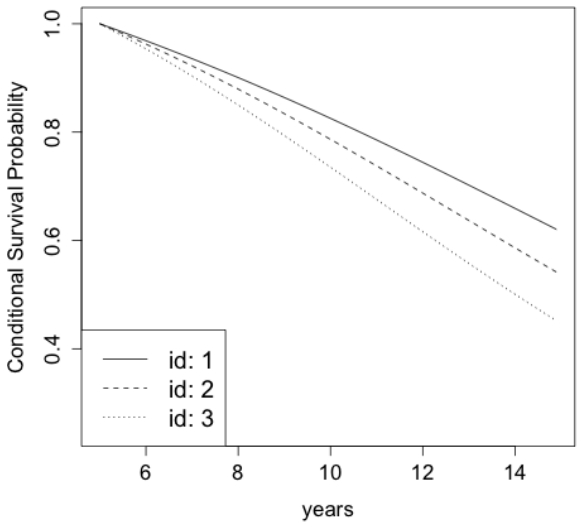

Figure 4: Estimated conditional progression-free probability of remaining years (after year 5) for the left eye, given the right eye has progressed by year 5, for subjects with different genotypes of *rs*2284665 (age 60 and severity score 3 in both eyes).

Likewise, we can also obtain the eye-level marginal survival probabilities (i.e., `plot_margin` $= 1$ for the left eyes) for the same three subjects, as illustrated in Figure 5.

```
plot(x = copula2_sp, class = "marginal", newdata = newdata2,
     plot_margin = 1, ylim = c(0.6,1), ylab = "Marginal Survival Probability")
```

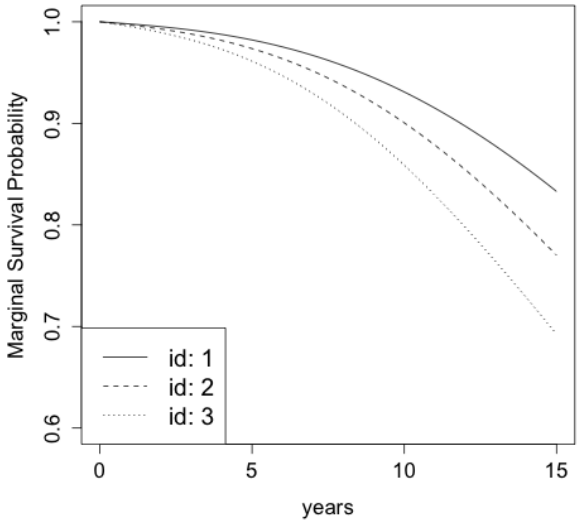

Figure 5: Estimated marginal progression-free probability for one eye from subjects with different genotypes of *rs*2284665 (age 60 and severity score 3 in both eyes).

Summary

This paper presents the R package **CopulaCenR** for implementing copula-based regression models in bivariate censored data, including both bivariate right-censored data and bivariate interval-censored data. A variety of Archimedean copulas, including a flexible two-parameter copula, are built in the

package to accommodate different dependence structures. Moreover, the package can fit various parametric and semiparametric regression models for the two margins within the copula function. In particular, a general semiparametric transformation model with PH and PO models being its special cases is implemented for the margins in this package. For parameter estimation, a novel two-step procedure is adopted to guarantee stable and fast computation. For the inference of covariate effects, all three likelihood-based tests are provided. Lastly, two real data examples are given to demonstrate the key features and capabilities of this package.

One future extension of this package is to allow multivariate copula functions for handling multivariate censored events. Another important research extension is to add goodness-of-fit tests, which is critical for choosing a proper copula model. However, there are limited works in testing copula models in bivariate censored data, especially in bivariate interval-censored data under the regression setting. The current literature (e.g., Shih, 1998; Andersen et al., 2010; Emura et al., 2010; Wang, 2010) only focus on testing copulas in bivariate right-censored data without covariates. We are currently investigating these directions and plan to incorporate them in a future version of **CopulaCenR**.

Acknowledgments

The Authors are grateful to Dr. Wei Chen for providing valuable suggestions about package development and the AREDS data analysis.

Bibliography

K. Aas, C. Czado, A. Frigessi, and H. Bakken. Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics*, 44(2):182 – 198, 2009. URL https://doi.org/10.1016/j.insmatheco.2007.02.001. [p267]

M. M. Ali, N. N. Mikhail, and M. S. Haq. A class of bivariate distributions including the bivariate logistic. *Journal of Multivariate Analysis*, 8(3):405–412, 1978. URL https://doi.org/10.1016/0047-259X(78)90063-5. [p268]

P. K. Andersen, C. T. Ekstrom, J. P. Klein, Y. Shu, and M.-J. Zhang. A class of goodness of fit tests for a copula based on bivariate right-censored data. *Biometrical Journal*, 47(6):815–824, 2010. URL https://doi.org/10.1002/bimj.200410163. [p279]

AREDS Group. The Age-Related Eye Disease Study (AREDS): Design implications. AREDS report no. 1. *Controlled Clinical Trials*, 20(6):573–600, 1999. URL https://doi.org/10.1016/s0197-2456(99)00031-8. [p266, 275]

N. E. Breslow. Discussion of the paper by D. R. Cox. *Journal of the Royal Statistical Society: Series B*, 34: 216–217, 1972. URL https://doi.org/10.2307/1403236. [p271]

D. G. Clayton. A model for association in bivariate life tables and application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65(1):141–151, 1978. URL https://doi.org/10.2307/2335289. [p266, 268]

D. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, 34(2):187–220, 1972. URL https://www.jstor.org/stable/2985181. [p267]

D. R. Cox and D. V. Hinkley. *Theoretical Statistics*. Chapman and Hall/CRC, 1979. URL https://doi.org/10.1201/b14832. [p272]

M. C. Donohue and R. Xu. **phmm**: *Proportional Hazards Mixed-effects Models*, 2019. URL https://CRAN.R-project.org/package=phmm. R package version 0.7-11. [p267]

T. Emura. **Copula.surv**: *Association Analysis of Bivariate Survival Data Based on Copulas*, 2018. URL https://CRAN.R-project.org/package=Copula.surv. R package version 1.0. [p267]

T. Emura, C. W. Lin, and W. Wang. A goodness-of-fit test for archimedean copula models in the presence of right censoring. *Computational Statistics & Data Analysis*, 54(12):3033–3043, 2010. URL https://doi.org/10.1016/j.csda.2010.03.013. [p267, 279]

M. J. Frank. On the simultaneous associativity of f (x , y) and x + y - f (x , y). *Aequationes Mathematicae*, 19(1):194–226, 1979. URL https://doi.org/10.1007/BF02189866. [p268]

E. J. Gumbel. Bivariate exponential distributions. *Journal of the American Statistical Association*, 55(292): 698–707, 1960. URL https://doi.org/10.2307/2281591. [p268]

M. Hofert, I. Kojadinovic, M. Maechler, and J. Yan. **copula**: *Multivariate Dependence with Copulas*, 2018. URL https://CRAN.R-project.org/package=copula. R package version 0.999-19. [p267]

P. Hougaard. *Analysis of Multivariate Survival Data*. Springer-Verlag, New York, 2000. URL https://doi.org/10.1002/sim.938. [p266]

W. J. Huster, R. Brookmeyer, and S. G. Self. Modelling paired survival data with covariates. *Biometrics*, 45(1):145–156, 1989. URL https://doi.org/10.2307/2532041. [p266, 273, 274]

H. Joe. Parametric families of multivariate distributions with given margins. *Journal of Multivariate Analysis*, 46(2):262–282, 1993. URL https://doi.org/10.1006/jmva.1993.1061. [p268]

H. Joe. *Multivariate Models and Multivariate Dependence Concepts*. Chapman & Hall, London, 1997. URL https://doi.org/10.1201/9780367803896. [p268]

H. Joe. *Dependence modeling with copulas*. CRC press, 2014. URL https://doi.org/10.1201/b17116. [p268]

H. Joe and T. Hu. Multivariate distributions from mixtures of max-infinitely divisible distributions. *Journal of multivariate analysis*, 57(2):240–265, 1996. URL https://doi.org/10.1006/jmva.1996.0032. [p268]

I. Kojadinovic and J. Yan. Modeling multivariate distributions with continuous margins using the copula R package. *Journal of Statistical Software*, 34(9):1–20, 2010. URL https://doi.org/10.18637/jss.v034.i09. [p267]

G. Lindfield et al. *Microcomputers in Numerical Analysis*. Halsted Press, 1989. URL https://doi.org/10.1112/S002557930001319X. [p272]

G. Marra and R. Radice. Bivariate copula additive models for location, scale and shape. *Computational Statistics & Data Analysis*, 112:99–113, 2017. URL https://doi.org/10.1016/j.csda.2017.03.004. [p267]

G. Marra and R. Radice. Copula link-based additive models for right-censored event time data. *Journal of the American Statistical Association*, pages 1–20, 2019. URL https://doi.org/10.1080/01621459.2019.1593178. [p267]

G. Marra and R. Radice. **GJRM**: *Generalised Joint Regression Modelling*, 2020. URL https://CRAN.R-project.org/package=GJRM. R package version 0.2-2. [p267]

G. Marra, R. Radice, T. Bärnighausen, S. N. Wood, and M. E. McGovern. A simultaneous equation approach to estimating HIV prevalence with nonignorable missing responses. *Journal of the American Statistical Association*, 112(518):484–496, 2017. URL https://doi.org/10.1080/01621459.2016.1224713. [p267]

G. Masarotto and C. Varin. Gaussian copula regression in R. *Journal of Statistical Software*, 77(8):1–26, 2017. URL https://doi.org/10.18637/jss.v077.i08. [p267]

M. Munda, F. Rotolo, and C. Legrand. **parfm**: Parametric frailty models in R. *Journal of Statistical Software*, 51(11):1–20, 2012. URL https://doi.org/10.18637/jss.v051.i11. [p267]

T. Nagler and T. Vatter. **gamCopula**: *Generalized Additive Models for Bivariate Conditional Dependence Structures and Vine Copulas*, 2020. URL https://CRAN.R-project.org/package=gamCopula. R package version 0.0-7. [p267]

J. Nash. *Compact Numerical Methods for Computers: Linear Algebra and Function Minimisation*. Taylor & Francis, 1990. URL https://doi.org/10.2307/3314683. [p274]

R. B. Nelsen. *An Introduction to Copulas*. Springer-Verlag, New York, 2006. URL https://doi.org/10.1007/0-387-28678-0. [p268]

D. S. Nicole Kraemer. *Bivariate Copula Based Regression Models*, 2014. URL https://cran.r-project.org/web/packages/CopulaRegression/. R package version 0.1-5. [p267]

D. Oakes. A model for association in bivariate survival data. *Journal of the Royal Statistical Society: Series B*, 44(3):414–422, 1982. URL https://www.jstor.org/stable/2345500. [p266]

L. Prenen, R. Braekers, and L. Duchateau. Extending the Archimedean copula methodology to model multivariate survival data grouped in clusters of variable size. *Journal of the Royal Statistical Society: Series B*, 79(2):483–505, 2017a. URL https://doi.org/10.1111/rssb.12174. [p267]

L. Prenen, R. Braekers, L. Duchateau, and E. D. Troyer. **Sunclarco***: Survival Analysis using Copulas*, 2017b. URL https://CRAN.R-project.org/package=Sunclarco. R package version 1.0.0. [p267]

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019. URL http://www.R-project.org/. [p267]

V. Rondeau, Y. Marzroui, and J. Gonzalez. **frailtypack**: An R package for the analysis of correlated survival data with frailty models using penalized likelihood estimation or parametrical estimation. *Journal of Statistical Software*, 47(4):1–28, 2012. URL https://doi.org/10.18637/jss.v047.i04. [p267]

U. Schepsmeier, J. Stoeber, E. C. Brechmann, B. Graeler, T. Nagler, T. Erhardt, C. Almeida, A. Min, C. Czado, M. Hofmann, et al. **VineCopula***: Statistical Inference of Vine Copulas*, 2018. URL https://CRAN.R-project.org/package=VineCopula. R package version 2.1.8. [p267]

J. H. Shih. A goodness-of-fit test for association in a bivariate survival model. *Biometrika*, 85(1):189–200, 1998. URL https://doi.org/10.1093/biomet/85.1.189. [p279]

A. Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8:229–231, 1959. URL https://doi.org/10.12691/ijefm-3-2-3. [p270]

J. Sun. *The Statistical Analysis of Interval-Censored Failure Time Data*. Springer Science & Business Media, 2007. URL https://doi.org/10.1007/0-387-37119-2. [p266]

T. Sun and Y. Ding. Copula-based semiparametric regression method for bivariate data under general interval censoring. *Biostatistics*, 2019. URL https://doi.org/10.1093/biostatistics/kxz032. [p269, 271, 272]

T. Sun, Y. Liu, R. J. Cook, W. Chen, and Y. Ding. Copula-based score test for bivariate time-to-event data, with application to a genetic study of AMD progression. *Lifetime Data Analysis*, 25(3):546–568, 2019. URL https://doi.org/10.1007/s10985-018-09459-5. [p268, 271, 272]

A. Swaroop, E. Y. Chew, G. R. Abecasis, et al. Unraveling a multifactorial late-onset disease: from genetic susceptibility to disease mechanisms for Age-related Macular Degeneration. *Annual Review of Genomics and Human Genetics*, 10:19–43, 2009. URL https://doi.org/10.1146/annurev.genom.9.081307.164350. [p275]

T. M. Therneau. **coxme***: Mixed Effects Cox Models*, 2018a. URL https://CRAN.R-project.org/package=coxme. R package version 2.2-10. [p267]

T. M. Therneau. **survival***: A Package for Survival Analysis in S*, 2018b. URL https://CRAN.R-project.org/package=survival. R package version 2.43-3. [p267]

T. Vatter and V. Chavez-Demoulin. Generalized additive models for conditional dependence structures. *Journal of Multivariate Analysis*, 141:147–167, 2015. URL https://doi.org/10.1016/j.jmva.2015.07.003. [p267]

A. Wang. Goodness-of-fit tests for Archimedean copula models. *Statistica Sinica*, 20:441–453, 2010. URL https://www.jstor.org/stable/24309000. [p279]

L. J. Wei, D. Lin, and L. Weissfeld. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, 84(408):1065–1073, 1989. URL https://doi.org/10.2307/2290084. [p266]

J. Yan. Enjoy the joy of copulas: With a package copula. *Journal of Statistical Software*, 21(4):1–21, 2007. URL https://doi.org/10.18637/jss.v021.i04. [p267]

Q. Zhou, T. Hu, and J. Sun. A sieve semiparametric maximum likelihood approach for regression analysis of bivariate interval-censored failure time data. *Journal of the American Statistical Association*, 112(518):664–672, 2017. URL https://doi.org/10.1080/01621459.2016.1158113. [p271]

Tao Sun
School of Statistics
Renmin University of China

59 Zhongguancun Street
Beijing, China
Department of Biostatistics
University of Pittsburgh
Pittsburgh, U.S.A.
ORCiD: 0000-0003-4447-3005
tao.sun@pitt.edu

Ying Ding
Department of Biostatistics
University of Pittsburgh
130 De Soto Street
Pittsburgh, U.S.A.
ORCiD: 0000-0003-1352-1000
yingding@pitt.edu