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 Sundarar 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

\[ C_\eta(u, v) = \phi_\eta\left( \phi_\eta^{-1}(u) + \phi_\eta^{-1}(v) \right), \]

where \( u \) and \( v \) are two uniformly distributed margins; \( \phi_\eta \) is the generator function, which is a continuous, strictly decreasing and convex function; \( \phi_\eta^{-1} \) is the inverse of \( \phi_\eta \). One generator function uniquely
determines an Archimedean copula. The copula parameter $\eta$ has a one-to-one correspondence with the popular dependence measure Kendall’s $\tau$. Another property of the copula is the tail dependence (i.e., $\tau_L$ and $\tau_U$ 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 $\eta$. Other Archimedean copulas, such as Frank (Frank, 1979), Joe (Joe, 1997) 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

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

The two dependence parameters ($\alpha$ and $\kappa$) are explicitly connected to Kendall’s $\tau$ with $\tau = 1 - 2\alpha x / (2x + 1)$, 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.

![Figure 1](image_url)

**Table 1**: Summary of implemented Archimedean copula families.

<table>
<thead>
<tr>
<th>Family</th>
<th>Parameter Space</th>
<th>Generator $\Phi_k(t, \alpha, \beta) \in (0, \infty)$</th>
<th>Generator Inverse $\Phi^{-1}_k(s, \alpha, \beta) \in (0, 1)$</th>
<th>$\tau_L$</th>
<th>$\tau_U$</th>
<th>Kendall’s $\tau$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clayton</td>
<td>$\eta &gt; 0$</td>
<td>$(-1 + x/\eta + 1)$</td>
<td>$(-1 + x/\eta)$</td>
<td>0</td>
<td>0</td>
<td>1 $- 1/\eta$</td>
</tr>
<tr>
<td>Gumbel</td>
<td>$q \geq 1$</td>
<td>$\exp(-t^{1/q})$</td>
<td>$(\log x)^q$</td>
<td>0</td>
<td>0</td>
<td>1 $- 1/q$</td>
</tr>
<tr>
<td>Frank</td>
<td>$q \geq 0$</td>
<td>$t^{1/q} - 1/(1 + q(t^{1/q} - 1))$</td>
<td>$(\log(1 + q/(1 + q - 1)))/q^q$</td>
<td>0</td>
<td>0</td>
<td>1 $+ 4D(q)/q^q$</td>
</tr>
<tr>
<td>AMH</td>
<td>$\eta \in (0, 1)$</td>
<td>$1 - (1 - x^\eta)$</td>
<td>$(-1 + x^\eta)/\log(1 + x^\eta)$</td>
<td>0</td>
<td>0</td>
<td>$1 - 2(1 - \eta)^2/\log(1 + x^\eta) + \eta^2/\log^2(1 + x^\eta)$</td>
</tr>
<tr>
<td>Joe</td>
<td>$\eta \geq 0$</td>
<td>$\log(1 - (1 - x^\eta))/\eta$</td>
<td>$(\log(1 - x^\eta))/\eta$</td>
<td>0</td>
<td>0</td>
<td>$1 - 2(1 - \eta)^2/\log(1 + x^\eta) + \eta^2/\log^2(1 + x^\eta)$</td>
</tr>
<tr>
<td>Copula2</td>
<td>$\alpha, \kappa \in (0, 1), \alpha &gt; 0$</td>
<td>$(1 + \alpha x)/\alpha - 1/\kappa$</td>
<td>$(-1 + x^\eta)/\kappa$</td>
<td>2 $- \alpha$</td>
<td>2 $+ \alpha$</td>
<td>$1 - 2\alpha x / (2x + 1)$</td>
</tr>
</tbody>
</table>

$\tau_L$ and $\tau_U$ are the lower and upper tail dependence measures.

$D(q)$ is the Dohby function written as $D(q) = \frac{1}{\sqrt{\pi q}} \frac{1}{\sqrt{1 - q}}$.
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).

<table>
<thead>
<tr>
<th>Type</th>
<th>Models</th>
<th>Survival Distributions $S(t)$</th>
<th>Corresponding R Functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parametric</td>
<td>Weibull</td>
<td>$\exp\left{-\frac{(t/\lambda)^{k}}{\lambda}e^{Z^\top \beta}\right}$</td>
<td>rc_par_copula, ic_par_copula</td>
</tr>
<tr>
<td></td>
<td>Gompertz</td>
<td>$\exp\left{-\frac{b}{a}(e^{at} - 1)e^{Z^\top \beta}\right}$</td>
<td>ic_par_copula</td>
</tr>
<tr>
<td></td>
<td>Loglogistic</td>
<td>$\left{1 + \frac{(t/\lambda)^{k}}{\lambda}e^{Z^\top \beta}\right}^{-1}$</td>
<td>ic_par_copula</td>
</tr>
<tr>
<td>Semiparametric</td>
<td>Cox Transformation</td>
<td>$\exp\left{-\Lambda(t)e^{Z^\top \beta}\right}$</td>
<td>ic_spCox_copula</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ic_spTran_copula</td>
</tr>
</tbody>
</table>

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 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 $\tau$ 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 $T_{ij}$, $\Delta_{ij}$ 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 $C_\eta$ 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 $C_\eta$ is called a copula and its parameter $\eta$ 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 $\theta = (\beta^\top, \delta^\top, \eta, S_{01}, S_{02})^\top$ all the unknown parameters in $S(t_1, t_2)$, where $\beta$ is the regression coefficient vector and $S_{0j}$ is the baseline survival function for the $j$th margin. Then, the joint likelihood for the observed data $D = \{D_i\}_{i=1}^n$ can be written as

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

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

\[
L_n(\theta | D) = \prod_{i=1}^n \Pr(L_{i1} < T_{i1} \leq R_{i1}, L_{i2} < T_{i2} \leq R_{i2}|Z_{i1}, Z_{i2}) \\
= \prod_{i=1}^n \left[ \Pr(T_{i1} > L_{i1}, T_{i2} > L_{i2}|Z_{i1}, Z_{i2}) - \Pr(T_{i1} > L_{i1}, T_{i2} > R_{i2}|Z_{i1}, Z_{i2}) \\
- \Pr(T_{i1} > R_{i1}, T_{i2} > L_{i2}|Z_{i1}, Z_{i2}) + \Pr(T_{i1} > R_{i1}, T_{i2} > R_{i2}|Z_{i1}, Z_{i2}) \right] \\
= \prod_{i=1}^n \left[ C_\eta(S_1(L_{i1}|Z_{i1}), S_2(L_{i2}|Z_{i2})) - C_\eta(S_1(L_{i1}|Z_{i1}), S_2(R_{i2}|Z_{i2})) \\
- C_\eta(S_1(R_{i1}|Z_{i1}), S_2(L_{i2}|Z_{i2})) + C_\eta(S_1(R_{i1}|Z_{i1}), S_2(R_{i2}|Z_{i2}) \right].
\]

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 $\infty$) or two (if one $R_{ij}$ is $\infty$) 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

$$S_j(t_j|Z_j) = \exp\{- (t_j / \lambda_j)^{b_j} e^{Z_j^\top \beta_j}\}, \quad j = 1, 2,$$
where \( \lambda_i \) and \( k_j \) are the scale and shape parameters of the baseline Weibull distribution, and \( \beta_j \) are the covariate effects. The model follows the PH assumption. In this case, the parameter set \( \theta \) becomes \((\beta_1^\top, \eta, \lambda_1, k_1, \lambda_2, k_2)^\top\). 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

\[
S_j(t_j|Z_j) = \exp\{-\Lambda_j(t_j)e^{Z_j\beta_j}\}, \quad j = 1, 2,
\]
in which the Breslow baseline cumulative hazard function \( \Lambda_j(t) \) is given by

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

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

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

\[
B_k(t, m, l, u) = \binom{m}{k} \frac{t-l}{u-l}(1-\frac{t-l}{u-l})^{m-k}, \quad k = 0, ..., m,
\]
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):

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

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 \( \exp\{-\Lambda_j(t)e^{Z\beta_j}\} \), which is essentially a PH model. Likewise, when \( G_j(x) = \log(1+x) \) at \( r = 3 \), \( S_j(t|Z) \) becomes \( \{1 + \Lambda_j(t)e^{Z\beta_j}\}^{-1} \), 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 \( \theta \). 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 \( \ell_m(\theta|D) = \sum_{j=1}^m \log L(\theta|D_j) \). 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 \( \theta_0 \):
   (a) \( (\hat{\beta}_j^{(1)}, \hat{s}_j^{(1)}) = \arg\max l_m(\beta_j, S_0), \) where \( l_m \) denotes the log-likelihood for the marginal model, \( j = 1, 2; \)

The R Journal Vol. 12/1, June 2020 ISSN 2073-4859
We now separate the arguments $n$, $\beta$, and $\eta$ into two parts: $\beta_m$ and $\beta_g$, where $\beta_m$ is the parameter set of interest for hypothesis testing and $\beta_g$ denotes the rest of the regression coefficients. In certain cases, $\beta_g$ can be the entire regression parameter $\beta$. 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.

Likelihood-based tests for covariate effects

We now separate $\beta$ into two parts: $\beta_m$ and $\beta_g$, where $\beta_m$ is the parameter set of interest for hypothesis testing and $\beta_g$ denotes the rest of the regression coefficients. In certain cases, $\beta_g$ can be the entire regression parameter $\beta$. 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 $\lambda$ (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 based on a specified copula model, marginal distributions and covariate values.

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.

```R
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
```
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

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

<table>
<thead>
<tr>
<th>id</th>
<th>ind</th>
<th>obs_time</th>
<th>status</th>
<th>treat</th>
<th>age</th>
<th>type</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>1</td>
<td>46.23</td>
<td>0</td>
<td>0</td>
<td>28</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>46.23</td>
<td>0</td>
<td>2</td>
<td>28</td>
<td>2</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>42.50</td>
<td>0</td>
<td>2</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>2</td>
<td>31.30</td>
<td>1</td>
<td>0</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>42.27</td>
<td>0</td>
<td>1</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>2</td>
<td>42.27</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>1</td>
</tr>
</tbody>
</table>

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`).

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

<table>
<thead>
<tr>
<th>estimate</th>
<th>SE</th>
<th>stat</th>
<th>pvalue</th>
</tr>
</thead>
<tbody>
<tr>
<td>lambda</td>
<td>90.64</td>
<td>13.18</td>
<td>47.24</td>
</tr>
<tr>
<td>k</td>
<td>0.81</td>
<td>0.06</td>
<td>189.10</td>
</tr>
<tr>
<td>treat1</td>
<td>-0.57</td>
<td>0.19</td>
<td>8.18</td>
</tr>
<tr>
<td>treat2</td>
<td>0.00</td>
<td>0.17</td>
<td>0.00</td>
</tr>
<tr>
<td>eta</td>
<td>0.62</td>
<td>0.26</td>
<td>5.65</td>
</tr>
</tbody>
</table>

Accounting for the significance of each coefficient is 0

Final llk: -839.7212
Convergence 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 $-k \log(\lambda) + \text{treat1} = -4.20$, and similarly the argon treatment effect is $-k \log(\lambda) + \text{treat2} = -3.63$, 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`.

```r
AIC(clayton_wb)
1689.442

BIC(clayton_wb)
1705.858
```

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

```r
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.

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

id    lp1    lp2
5     0.00000000 0.005299655
14    0.005299655 0.00000000
16    0.571449035 0.00000000

When type is "survival", the fitted outputs are marginal ($S_1, S_2$) and joint ($S_{12}$) survival probabilities at the observed times ($t_1, t_2$).
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 0.4655284

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 newdata 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)))

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
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`).

```r
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 0.0426100 0.012271 12.057 0.0005159 ***
rs2284665 0.3977120 0.091180 19.026 1.290e-05 ***
SevScaleBL 0.7226810 0.053258 184.132 < 2.2e-16 ***
alpha 0.9305080 0.058714 251.136 < 2.2e-16 ***
kappa 0.9740370 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 rs2284665 is \( \exp(0.397712) = 1.49 \) with a p-value \( 1.29 \times 10^{-5} \), 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.

```r
AIC(copula2_sp)
4226.356
BIC(copula2_sp)
4266.353
```

Also, the estimated Kendall’s \( \tau \) is 0.38, suggesting moderate dependence in AMD progression between two eyes.

```r
tau_copula(eta = as.numeric(coef(copula2_sp)[c("alpha","kappa")]),
    copula = "Copula2")
0.3851248
```
Furthermore, we can test the effect of rs2284665 by the generalized score test. We first fit a null model without rs2284665 and then test its effect using the function score_copula.

```r
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.

```r
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.

```r
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
2  1   3      60       1
3  1   3      60       2
4  1   3      60       0
5  1   3      60       1
6  1   3      60       2
```

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

![Figure 3: Estimated joint progression-free probability contours for subjects with different genotypes of rs2284665 (age 60 and severity score 3 in both eyes).](image)

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).

```r
plot(x = copula2_sp, class = "conditional", newdata = newdata2,
     cond_margin = 2, cond_time = 5, ylim = c(0,25),
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 rs2284665 (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 rs2284665 (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


Tao Sun
School of Statistics
Renmin University of China