# Analysis of Bivariate Survival Data based on copulas with logGEV marginals

Dooti Roy

February 18, 2015

**Abstract**

This project introduces a novel copula based methodology to analyze right censored bivariate survival data using flexible log-GEV marginals. The Clayton Copula structure is used to represent the dependency between the survival times. In the first section, a brief introduction to the problem and current methods are outlined including their roadblocks and restrictions. In the second section, our method is described. In the third section some simulation results are shown. We conclude with future work directions and limitations.

## 1 Introduction

During the last two decades, there have been a growing interest in modeling multivariate survival data. It has been medically observed that incidence of one disease often increases the risk of another in a patient as in HIV patients. In Huster et al. (1989) the authors had introduced a fully parametric approach to analyze bivariate paired data with covariates. In Sahu and Dey (2000), the authors have used a frailty model to model bivariate survival data. Although these methods are fairly attractive since they are mostly identifiable and produces smooth survival functions, they have one serious drawback in terms of computational efficiency. Similarly the fully non parametric approach provides flexibility but is inefficient and in the presence of censoring, can be inconsistent as well.

Several researchers have studied bivariate distributions in copula models, including Clayton (1978) and Nelson (1986). Shih and Louis (1995) used a semi parametric copula based model to analyze bivariate survival data. One useful feature of the copula models is that the univariate marginals do not depend on the choice of the dependency structure and can be consequently be estimated separately than the dependency. In this paper, the goal is to develop the Bayesian framework for inference and estimation of the marginals and the dependence parameter following the approach outlined in Roy (2014) using Clayton copula structure and the flexible GEV distribution to model the marginal distributions.

## 2 Marginal Distributions and Bivariate Survival Model

### 2.1 Generalized Extreme Value Distribution as Marginal

We start by defining the generalized extreme value distribution following Roy et al. (2013). Going forward, the notations MGEV will be used as in Roy et al. (2013) for representing the GEV maxima model. If we assume a MGEV distribution for log T, where T is a marginal survival time i.e., log T $\sim$ MGEV$(\mu, \sigma, \xi)$ , then the corresponding pdf and survival function for T, i.e. T $\sim$ log MGEV$(\mu, \sigma, \xi)$ are respectively

$$f_M(t|\mu,\sigma,\xi) = \begin{cases} \frac{\exp[-(1+\xi\frac{\log t-\mu}{\sigma})^{-\frac{1}{\xi}}]}{\sigma t(1+\xi\frac{\log t-\mu}{\sigma})^{\frac{1}{\xi}+1}} & t > \exp(\mu - \frac{\sigma}{\xi}) \text{ if } \xi > 0 \textit{ or} \\ & t < \exp(\mu - \frac{\sigma}{\xi}) \text{ if } \xi < 0 \\ \frac{1}{\sigma t}\exp(-\frac{\log t-\mu}{\sigma})\exp[-\exp(-\frac{\log t-\mu}{\sigma})] & 0 < t < \infty \text{ if } \xi = 0 \end{cases}$$

and the survival function which is $(1 - F_M(t|\mu,\sigma,\xi))$ is given by,

$$S_M(t|\mu,\sigma,\xi) = \begin{cases} 1 - \exp[-(1+\xi\frac{\log t-\mu}{\sigma})^{-\frac{1}{\xi}}] & \text{if } \xi \neq 0 \\ 1 - \exp(-\exp(-\frac{\log t-\mu}{\sigma})) & \text{if } \xi = 0 \end{cases}$$

### 2.2 Bivariate Model

**Model:**

Let $(T_1, T_2)$ denote failure times of two events for each subject or failure times of members of each group. Marginally we assume that, $\log T_1 \sim$ MGEV $(\mu_1, \sigma_1, \xi_1)$ and $\log T_2 \sim$ MGEV $(\mu_2, \sigma_2, \xi_2)$ for $i = 1, 2$, where MGEV is defined above following Roy et al. (2013). The joint survival function based on a copula $C_\phi, \phi \in G$ is given by

$$S(t_1, t_2) = C_\phi(S_1(t_1), S_2(t_2)),$$

where $S_1(t_1)$ and $S_2(t_2)$ are the marginal survival functions obtained from the previous section. The parameter $\phi$ measures the "intensity" of dependence between the individual failure times. We use the popular bivariate copula from the Clayton Family:

$$C_\phi(u_1, u_2) = max((u_1^{-\phi} + u_2^{-\phi} - 1)^{-1/\phi}, 0)$$

In this case $G = (-1, \infty) \setminus \{0\}$.

### 2.3 Model Settings

Let $T_{ij}(C_{ij})$ be the survival (censoring) time of the $j$th component for the $i$th subject, $j = 1, 2; i = 1, 2, ..., n.$. We assume $(T_{i1}, T_{i2})$ and $(C_{i1}, C_{i2})$ are independent and $(T_{i1}, T_{i2})$, $i = 1, 2, ..., n.$ be *iid* with common pdf $f(t_1, t_2)$ and survival time $S(t_1, t_2)$. So the observed data is $(\boldsymbol{y_1, y_2, \delta_1, \delta_2})$ where $\boldsymbol{y_i} = (y_{1i}, y_{2i}, ..., y_{ni})$, $\boldsymbol{\delta_i} = (\delta_{1i}, \delta_{2i}, ..., \delta_{ni})$. Let $y_{ij} = min(T_{ij}, (C_{ij}))$ and $\delta_{ij} = I(y_{ij} = T_{ij})$. Let $\boldsymbol{y} = (\boldsymbol{y_1, y_2})$ and $\boldsymbol{\delta} = (\boldsymbol{\delta_1, \delta_2})$. Here $(S_1, S_2)$ and $(f_1, f_2)$ are the marginal survival and density functions of logGEV respectively as defined before. $\theta_1$ and $\theta_2$ are the parameters associated with each of the marginals.

Then the complete data likelihood function of the parameters $(\boldsymbol{\theta}, \xi)$ can be written as:

$$\begin{aligned}
L(\theta_1, \theta_2, \phi | \boldsymbol{y, \delta}) &= \prod_{i=1}^{n} (f(y_{i1}, y_{i2}))^{\delta_{i1}\delta_{i2}} \Big(\frac{\partial S(y_{i1}, y_{i2})}{\partial y_{i1}}\Big)^{\delta_{i1}(1-\delta_{i2})} \Big(\frac{\partial S(y_{i1}, y_i 2)}{\partial y_{i2}}\Big)^{\delta_{i2}(1-\delta_{i1})} \\
&\quad (S(y_{i1}, y_{i2}))^{(1-\delta_{i1})(1-\delta_{i2})} \\
&= \prod_{i=1}^{n} (c_\phi(S_{1\theta_1}(y_{i1}), S_{2\theta_2}(y_{i2}) f_{1\theta_1}(y_{i1}), f_{2\theta_2}(y_{i2})^{\delta_{i1}\delta_{i2}} \\
&\quad \Big( - \frac{\partial C_\phi(S_{1\theta_1}(y_{i1}), S_{2\theta_2}(y_{i2})}{\partial S_{1\theta_1}(y_{i1})} (-f_{1\theta_1}(y_{i1}))\Big)^{\delta_{i1}(1-\delta_{i2})} \\
&\quad \Big( - \frac{\partial C_\phi(S_{1\theta_1}(y_{i1}), S_{2\theta_2}(y_{i2})}{\partial S_{2\theta_2}(y_{i1})} (-f_{2\theta_2}(y_{i2}))\Big)^{\delta_{i2}(1-\delta_{i1})} \\
&\quad (C_\phi(S_{1\theta_1}(y_{i1}), S_{2\theta_2}(y_{i2}))^{(1-\delta_{i1})(1-\delta_{i2})}
\end{aligned} \tag{1}$$

## 3 Proposed Methodology

### 3.1 Stage I

There is mainly two developed approach to solve the problem. Shih and Louis (1995) proposed a two step procedure. First assume independence between failure times and estimate the marginals and then estimate $\phi$ assuming the estimated marginals are fixed. The Bayesian approach simultaneously estimates $(\theta_1, \theta_2, \phi)$ from joint posterior. The issue with the first approach is unnatural assumption of independence to begin with. In the second case finding appropriate prior of $\phi$ is difficult as it depends on which copula family is being used.

We start start by putting appropriate priors on $\theta_1$ and $\theta_2$. For fixed $\phi$, the joint posterior density of $\theta_1$ and $\theta_2$ is given by

$$\pi_\phi(\theta_1, \theta_2|\boldsymbol{y}, \boldsymbol{\delta}) = \frac{L(\theta_1, \theta_2, \phi|\boldsymbol{y}, \boldsymbol{\delta})\pi(\theta_1)\pi(\theta_2)}{m_\phi(\boldsymbol{y}, \boldsymbol{\delta})}, \;\; (\theta_1, \theta_2) \in \Theta \tag{2}$$

where $m_\phi(\boldsymbol{y}, \boldsymbol{\delta})$ is the normalizing constant given by

$$m_\phi(\boldsymbol{y}, \boldsymbol{\delta}) = \int_\Theta L(\theta_1, \theta_2, \phi|\boldsymbol{y}, \boldsymbol{\delta})\pi(\theta_1)\pi(\theta_2)d\theta_1 d\theta_2.$$

Following Roy (2014), we select that value of $\phi \in G$ which maximizes the marginal likelihood of the data $m_\phi(\boldsymbol{y}, \boldsymbol{\delta})$. Instead of maximizing $m_\phi(\boldsymbol{y}, \boldsymbol{\delta})$, we choose to maximize, $a \times m_\phi(\boldsymbol{y}, \boldsymbol{\delta})$ as it often easier. We choose "$a$" as $m_{\phi_1}(\boldsymbol{y}, \boldsymbol{\delta})$ for a pre fixed value $\phi_1$. Then by ergodic theorem we have a simple consistent estimator of $B_{\phi,\phi_1}$,

$$\frac{1}{N}\sum_{l=1}^{N} \frac{L(\theta_1^{(l)}, \theta_2^{(l)}, \phi|\boldsymbol{y}, \boldsymbol{\delta})}{L(\theta_1^{(l)}, \theta_2^{(l)}, \phi_1|\boldsymbol{y}, \boldsymbol{\delta})} \overset{a.s.}{\to} \int_\Theta \frac{L(\theta_1, \theta_2, \phi|\boldsymbol{y}, \boldsymbol{\delta})}{L(\theta_1, \theta_2, \phi|\boldsymbol{y}, \boldsymbol{\delta})}\pi_{\phi_1}(\theta_1, \theta_2|\boldsymbol{y}, \boldsymbol{\delta})d\theta_1 d\theta_2 = \frac{m_\phi(\boldsymbol{y}, \boldsymbol{\delta})}{m_{\phi_1}(\boldsymbol{y}, \boldsymbol{\delta})}, \tag{3}$$

as $N \to \infty$ where $\{\theta_1^{(l)}, \theta_2^{(l)}\}_{l=1}^N$ is a single Harris ergodic Markov chain with stationary density $\pi_{\phi_1}(\theta_1, \theta_2|\boldsymbol{y}, \boldsymbol{\delta})$.

Once the association parameter is determined, Gibb's sampler was used to determine the marginal parameter estimates.

## 3.2 Stage II

Roy (2014) mentioned that the above estimate of $\phi$ although simple is often unstable. To remove the instability introduced due to an arbitrary choice of $\phi_1$, (Roy, 2014) proposed a revised method. Let $\phi_1, \phi_2, \ldots, \phi_k \in G$ be $k$ appropriately chosen skeleton points. Let $\{\theta_1^{(j;l)}, \theta_2^{(j;l)}\}_{l=1}^{N_j}$ be a Markov chain with stationary density $\pi_{\phi_j}(\theta_1, \theta_2|\boldsymbol{y}, \boldsymbol{\delta})$ for $j = 1 \ldots, k$. Define $r_i = m_{\phi_i}(\boldsymbol{y}, \boldsymbol{\delta})/m_{\phi_1}(\boldsymbol{y}, \boldsymbol{\delta})$ for $i = 2, 3, \ldots, k$, with $r_1 = 1$. Then $B_{\phi,\phi_1}$ is consistently estimated by

$$\hat{B}_{\phi,\phi_1} = \sum_{j=1}^{k}\sum_{l=1}^{N_j} \frac{L(\theta_1^{(j;l)}, \theta_2^{(j;l)}, \phi|\boldsymbol{y}, \boldsymbol{\delta})}{\sum_{i=1}^{k} N_i L(\theta_1^{(j;l)}, \theta_2^{(j;l)}, \phi_i|\boldsymbol{y}, \boldsymbol{\delta})/\hat{r}_i}, \tag{4}$$

where $\hat{r}_1 = 1$, $\hat{r}_i$, $i = 2, 3, \ldots, k$ are consistent estimator of $r_i$'s obtained by the "reverse logistic regression" method proposed by (Geyer, 1994).

# 4 Simulation

## 4.1 Stage I

R package ”Copula” was used to generate bivariate $(u_1, u_2) \sim$ U(0,1). The data has dependency according to previously mentioned Clayton Copula structure. Next we used ”Inverse Survival Function” approach to get survival times generated from $mGEV(0, 1, \xi)$.

- Case I : $\xi \neq 0$.

$$\begin{aligned} 1 - u =& S(t) \\ =& \exp\,[-(1 + \xi \log(t))^{\frac{1}{\xi}+}] \\ \Rightarrow t =& \exp\,[\frac{-1 + (-\log(u))^{\xi}}{\xi}] \end{aligned} \tag{5}$$

- Case II: $\xi = 0$.

$$\begin{aligned} 1 - u =& S(t) \\ =& \exp\,(-t) \\ \Rightarrow t =& -\log(1 - u) \end{aligned} \tag{6}$$

Select $\xi_1 = \xi_2 = 0.3$ and $\phi = 5$. Censoring percentages in both $T_1$ and $T_2$ were taken to be about 12%. 3000 MCMC samples were generated and the first 1000 were discarded as burnin. A uniform prior of (-0.7, 0.7) was considered for $\xi$. The result of the simulation is displayed in table 1. *The accuracy of the results depends very much on the choice of the parameter $\phi_1$. As $\phi_1$ is chosen far from true $\phi$, the HPD intervals got wider and standard error of the estimates got larger. To control for the instability we implemented Stage II of the methodology following Roy (2014) and Geyer (1994).*

| $\phi_1$ | $Param$ | Estimate[S.E.] | 95% HPD |
|---|---|---|---|
| 1.00 | $\hat{\phi}$ | 5.2114 | - |
| | $\xi_1$ | 0.308[0.009] | (0.293, 0.323) |
| | $\xi_2$ | 0.312[0.009] | (0.299, 0.326) |
| 2.00 | $\hat{\phi}$ | 5.2116 | - |
| | $\xi_1$ | 0.308[0.008] | (0.292,0.322) |
| | $\xi_2$ | 0.312[0.006] | (0.300,0.324) |
| 4.00 | $\hat{\phi}$ | 5.2116 | - |
| | $\xi_1$ | 0.307[0.008] | (0.291,0.321) |
| | $\xi_2$ | 0.312[0.007] | (0.300,0.321) |
| 6.00 | $\hat{\phi}$ | 5.2115 | - |
| | $\xi_1$ | 0.308[0.008] | (0.291,0.322) |
| | $\xi_2$ | 0.310[0.009] | (0.294,0.322) |
| 8.00 | $\hat{\phi}$ | 5.2117 | - |
| | $\xi_1$ | 0.308[0.008] | (0.294,0.322) |
| | $\xi_2$ | 0.310[0.008] | (0.295,0.326) |

### 4.2 Stage II

We selected four skeletal points (2, 4, 6, 8). 1000 data points were simulated following similar method as in Stage I. This time the true value of $\phi$ was 5 and $\xi_1 = \xi_2 = 0.3$. For each of the four $\phi$, 3000 MCMC samples were generated. Using the MCMC samples, following Geyer (1994), first $\hat{r}_i$, $i = 2, 3, 4$ were estimated (we assume $hatr_1 = 1$ due to model identifiability issues) and then using them, estimate of $\phi$ was obtained. Then using $\phi$, $\xi_1$ and $\xi_2$ were estimated. Table 2 provides the details.

| $Param$ | Estimate[S.E.] | 95% HPD |
|---|---|---|
| $\hat{\phi}$ | 5.212 | - |
| $\xi_1$ | 0.[0.018] | (0.502,0.569) |
| $\xi_2$ | 0.540[0.015] | (0.510,0.562) |

## 5 Real Data Application

We consider the well known Diabetes Retinopathy Data for our analyses. This data was first introduced by Huster et al. (1989) in their seminal paper on modeling bivariate data with covariates. The Diabetes Retinopathy is the leading cause of blindness in Unites States under 60 years accounting for approximately 12% of all new cases. It has been shown that 90% of patients who have been diabetic for more than a decade will eventually develop retinopathy, or an ocular manifestation of blindness. Huster et al. (1989) analyzed the data using a frequentist approach with the Clayton family with cross-ratio function $\theta(\nu) = \alpha\ \alpha > 1$ and Weibull marginal distributions. Therneau and Grambsch (2000) considered random effect models and Sahu and Dey (2000) considered exponential and Weibull bivariate distributions with a Bayesian approach. J.S. and N.I. (2006) considered copulas to model the dependency of the data and follow the same two step procedure outlined by Shih and Louis (1995). The DRS data consists of 197 patients with severe Retinopathy affecting the eye.

# 6 Discussion

The method using multiple skeletal points is much more efficient. Results show that choosing skeletal points reasonably away from true value of $\phi$ still provided decent estimate. However it appears that we have delayed convergence in case one of the skeletal points is far away from true value. As next step, to conduct a real data analysis, the model must be expanded to incorporate covariates. Our future plan includes a real data analysis and model selection using multiple copula structures.

# References

Clayton, D. (1978), "A Model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease coincidence," *Biometrika*, 65, 141–151.

Geyer, C. J. (1994), "Estimating Normalizing Constants and reweighting Mixtures in Markov Chain Monte Carlo," *Technical Report.*

Huster, W., Brookmeyer, R., and Self, S. (1989), "Modelling Paired Survival Data with Covariates," *Biometrics*, 45, 145–156.

J.S., R. and N.I., T. (2006), "Bivariate survival modeling: a Bayesian approach based on copulas," *Lifetime Data Analysis*, 12, 205–222.

Nelson, R. (1986), "Properties of a one-parameter family of bivariate distributions with specified marginals," *Communications in Statistics, Part A*, 153, 3277–3285.

Roy, D., Roy, V., and Dey, D. (2013), "Analysis of survival data with a cure fraction under generalized extreme value distribution," *Tech. Report 49, Department of Statistics, University of Connecticut.*

Roy, V. (2014), "Efficient Estimation of the link function parameter in a robust bayesian binary regression model," *Computational Statistics and Data Analysis*, 73, 87–102.

Sahu, S. and Dey, D. (2000), "A Comparison of Frailty and Other Models for Bivariate Survival Data," *Lifetime Data Analysis*, 6, 207–228.

Shih, J. and Louis, T. (1995), "Inferences on the Association Parameter in Copula Models for Bivariate Survival Data," *Biometrics*, 51, 1384–1399.